Faithfulness in Abstractive Summarization: Progress and Challenges

Faisal Ladhak

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

Abstract

Faithfulness in Abstractive Summarization:

Progress and Challenges

Faisal Ladhak

The exponential increase in online text has created a pressing need for automatic summarization systems that can distill key information from lengthy documents. While neural abstractive summarizers have achieved gains in fluency and coherence, a critical challenge that has emerged is ensuring faithfulness, i.e., accurately preserving the meaning from the original text. Modern neural abstractive summarizers can distort or fabricate facts, undermining their reliability in real-world applications. Thus, this thesis tackles the critical issue of improving faithfulness in abstractive summarization.

This thesis is comprised of four parts. The first part examines challenges in evaluating summarization faithfulness, including issues with reference-free metrics and human evaluation. We propose a novel approach for building automated evaluation metrics that are less reliant on spurious correlations and demonstrate significantly improved performance over existing faithfulness evaluation metrics. We further introduce a novel evaluation framework that enables a more holistic assessment of faithfulness by accounting for the abstractiveness of summarization systems. This framework enables more rigorous faithfulness evaluation, differentiating between gains from increased extraction versus improved abstraction.

The second part focuses on explaining the root causes of faithfulness issues in modern summarization systems. We introduce a novel contrastive approach for attributing errors that vastly

outperforms prior work at tracing hallucinations in generated summaries back to training data deficiencies. Moreover, incorporating our method's ideas into an existing technique substantially boosts its performance. Through a case study, we also analyze pre-training biases and demonstrate their propagation to summarization models, yielding biased hallucinations. We show that while mitigation strategies during finetuning can reduce overall hallucination rates, the remaining hallucinations still closely reflect intrinsic pre-training biases.

The third part applies insights from previous sections to develop impactful techniques for improving faithfulness in practice. We propose a novel approach for adaptively determining the appropriate level of abstractiveness for a given input to improve overall faithfulness. Our method yields systems that are both more faithful and more abstractive compared to baseline systems. We further leverage our error attribution approach to clean noisy training data, significantly reducing faithfulness errors in generated outputs. Models trained on datasets cleaned with our approach generate markedly fewer hallucinations than both baseline systems and models trained using other data cleaning techniques.

Finally, the fourth part examines the summarization capabilities of LLMs and assesses their faithfulness. We demonstrate that instruction-tuning and RLHF are key for enabling LLMs to achieve high-quality zero-shot summarization in the news domain, with state-of-the-art LLMs generating summaries comparable to human-written ones. However, this ability does not extend to narrative summarization, where even advanced LLMs struggle to produce consistently faithful summaries. Finally, we highlight the difficulty in evaluating high-performing LLMs, showing that crowdsourcing evaluations of LLM outputs may no longer be reliable as fluency and coherence improve. We observe a substantial gap between crowd workers and experts in identifying deficiencies in LLM-generated narrative summaries.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest gratitude to my advisor, Kathleen McKeown, for her invaluable guidance, support, patience, and encouragement throughout my PhD journey. She struck the perfect balance of providing me the freedom to explore new research directions while steering me with insightful feedback. I consider myself incredibly fortunate to have had an advisor who was not only an academic role model to emulate but also understanding and caring on a personal level.

I would also like to thank Tatsunori Hashimoto and Dan Jurafsky, who enthusiastically embraced me into Stanford's NLP group and generously provided a workspace on campus so that I would not feel isolated during my remote stretches away. The opportunities for mentorship, collaboration, and growth they offered truly enriched my PhD journey. I would also like to thank Smaranda Muresan and Zhou Yu for serving on my dissertation committee and for their insightful feedback on my research.

I am profoundly grateful to the phenomenal collaborators and mentors I've worked with during my PhD journey. In particular, I express my deepest thanks to Esin Durmus, Claire Cardie, He He, Tianyi Zhang, Mirac Suzgun, Griffin Adams, and Melanie Subbiah for their invaluable contributions that made this dissertation possible. I would also like to acknowledge my friends and colleagues in the NLP group for the stimulating discussions over the years (please accept my apologies if I have forgotten anyone): Emily Allaway, Amith Ananthram, Tuhin Chakrabarty, Chris Hidey, Chris Kedzie, Vishwesh Kirthivasan, Fei-Tzin Lee, Samantha Mayers, Kailash Saravanakumar, Elsbeth Turcan, and David Wan.

I have been incredibly fortunate to be surrounded by amazing friends. I wish to especially thank Griffin Adams, Ilse Alejo, Deniz Altinbuken, Eugene Bagdasaryan, Nadia Bagdasaryan, Connie Ong Blukis, Valts Blukis, Rasool Fakoor, Shakil Farishta, Yiqing Hua, Ilia Ilmer, Daniel Lehman, Pablo Lujambio, Prashanth Mathur, Azade Nova, Kashif Shah, Mirac Suzgun, Julia Tolkacheva, Andreas Veit, and Taher Yusufali. I am eternally thankful for my friends, who have been pivotal in keeping me driven and uplifting me out of stressful moments. Their unwavering support and encouragement have truly made the completion of this journey possible.

I am eternally grateful to my parents for their tremendous personal sacrifices that have enabled me to reach this milestone. They courageously immigrated to a foreign land, having a minimal grasp of the language and limited social support, simply to provide increased opportunities for me. Despite adversity, they have raised and nurtured me with absolute love while working tirelessly to build a better life for our family. My achievements stand as a testament to their vision and dedication as parents. I am overwhelmed with gratitude for the unconditional support they have given me. I would also like to thank my sister, Farheen, who has been a constant source of encouragement, always willing to lend an ear when I needed someone to talk to during both the challenges and triumphs of pursuing this degree.

Finally, I would like to thank my wife, Esin, for her steadfast love and support throughout this long journey. I have been incredibly fortunate to have her by my side, celebrating the highs and uplifting me through the lows. Her keen intellect and collaborative spirit have contributed meaningfully to my research. But most of all, her unwavering belief in me made this dissertation possible. I could not have asked for a better life partner and friend by my side. Simply put, I would not be where I am today without her.

Dedication

For my beloved wife, whose endless love and unwavering support have always inspired me to pursue my dreams.

Chapter 1: Introduction and Background

As the amount of text data available online continues to grow exponentially, automatic text summarization techniques are increasingly needed to help people quickly understand the critical information contained in large documents or collections of documents. A summary is defined as a shortened piece of text that contains the most salient content from one or more longer input documents (K. S. Jones 1998; Radev, Hovy, et al. 2002). As such, text summarization involves two core tasks: i) extracting the most relevant content from the input, and ii) presenting the extracted content in a concise and coherent manner (K. S. Jones 1998).

Automatically generating high-quality summaries requires building systems that can analyze the discourse structure and meaning of the input text in order to identify and synthesize salient information in a manner that helps the reader efficiently grasp the key points. Early work on automatic summarization focused primarily on extractive methods, where the primary goal was to determine which sentences from the original text were most important to include in the summary (Salton et al. 1997; Marcu 1997; Barzilay, Kathleen R. McKeown, and Elhadad 1999; Mani and Bloedorn 1999; Radev, Jing, et al. 2000). While purely extractive methods address the key challenge of selecting important content, the resulting summaries often lack clarity, structure, and coherence, as they consist of extracted sentences joined together without any rewrite (Dang 2005; Nenkova et al. 2011).

Prior work has shown that humans rarely copy entire sentences when writing summaries – instead, they rely on more abstractive operations, such as paraphrasing and fusion, in order to produce more fluent, coherent summaries (Jing and Kathleen R McKeown 1999). This has led to increased research toward building abstractive summarization systems. Initial approaches to building abstractive summarizers focused on post-processing extractive summaries using specific operations such as sentence fusion (Barzilay and Kathleen R. McKeown 2005; Filippova and Strube

2008; Thadani and K. McKeown 2013) and sentence compression (Jing 2000; Knight and Marcu 2002; McDonald 2006; Cohn and Lapata 2008).

Source Article: People who are addicted to a drug are especially driven to find loopholes. In countries where marijuana has already been introduced for medical purposes, this has been the case. Legalizing marijuana would pose a bad example and trigger pressure for the legalization of other drugs.
Generated Summary: Marijuana should not be made legal for recreational use, but it should be legal for medical use, according to the World Health Organization.

Table 1.1: An example of an unfaithful summary generated by the BART-XSum model.

The development of large-scale datasets (Rush et al. 2015; Hermann et al. 2015) and improved neural sequence-to-sequence models (Sutskever et al. 2014; Bahdanau et al. 2015; Vaswani et al. 2017) enabled more data-driven, end-to-end training of abstractive summarization systems. This led to steady progress in improving the fluency of generated summaries (See et al. 2017; Y.-C. Chen and Bansal 2018; L. Dong et al. 2019). Recently, pretrained language models have catalyzed a major leap forward in building effective abstractive summarization systems. Summarizers built on top of pretrained language models are capable of producing remarkably fluent and human-like abstractive summaries (Yang Liu and Lapata 2019; J. Zhang et al. 2020; Lewis et al. 2020).

However, despite the tremendous progress, current summarization systems still face a critical limitation - they can sometimes generate unfaithful summaries that do not accurately reflect the information in the source text (Z. Cao et al. 2018). Given their remarkable fluency, unfaithful summaries generated by current summarizers can appear very plausible and, therefore, could potentially contribute to the spread of misinformation. Table 1.1 shows an example of how a state-of-the-art abstractive summarization system can potentially spread misinformation by generating plausible but fabricated summaries. In this instance, not only does the summarizer generate a claim that is opposite of what was stated in the article, but it also attributes this claim to a credible organization not mentioned in the original article. Given this concern, evaluating and improving the faithfulness of abstractive summarization systems has emerged as a critical research direction in the field, as the faithfulness issue needs to be solved in order to build more reliable, trustworthy summarization systems that can be deployed in real-world settings. Therefore, this thesis tackles the critical issue of faithfulness in abstractive summarization with the goal of understanding the main causes of the problem and proposing potential mitigations towards building more reliable abstractive summarization systems.

While this thesis focuses on faithfulness in abstractive summarization, we note that extractive summarizers may also make faithfulness errors. Early extractive work was presumed to have fewer faithfulness issues since summaries were composed mainly of extracted sentences with simple compression or fusion. The faithfulness of these systems, however, was not explicitly measured – recent studies show extractive methods still make faithfulness mistakes, albeit different kinds than abstractive systems (Shiyue Zhang et al. 2022).

1.1 Overview of Thesis

In this thesis, we dive into the faithfulness issue in abstractive summarization and provide contributions in four key areas: 1) the evaluation of faithfulness, 2) attribution for faithfulness errors, 3) approaches for mitigating faithfulness errors, and 4) assessing faithfulness in LLMs.

In chapter 2, we examine the current approaches to evaluating summarization systems, focusing on assessing the faithfulness of generated summaries. Progress in the field of abstractive summarization has primarily been measured by comparing model-generated summaries against reference summaries using automated evaluation metrics such as ROUGE (Lin 2004). However, these metrics have been shown to correlate poorly with human assessments of faithfulness (Durmus, H. He, et al. 2020; A. Wang et al. 2020; Maynez et al. 2020), necessitating the development of new metrics geared towards measuring faithfulness. Recent work has proposed reference-free faithfulness metrics aimed at addressing these limitations (Durmus, H. He, et al. 2020; Kryscinski, McCann, et al. 2020; Goyal and Greg Durrett 2020). However, in §2.1, we critically analyze recently proposed evaluation metrics to understand whether they can be used to evaluate faithfulness and show that the effectiveness of these metrics may be overstated. We show that they primarily exploit the spurious correlation between extractiveness and faithfulness rather than capturing the semantics of what it means for a summary to be faithful. We then propose a new metric that addresses this spuriousness issue and show that it outperforms prior work.

Beyond evaluation metrics, the correlation between extractiveness and faithfulness complicates comparative assessments of summarization systems, even when using humans to evaluate. In §2.2, we show that naively comparing faithfulness scores of systems can lead to misleading conclusions, potentially at the detriment of building better abstractive summarization systems. We propose a novel evaluation framework that evaluates faithfulness more holistically by taking into account the abstractiveness of different summarization systems. Using our new framework, we show that recently proposed methods aimed at improving the faithfulness of summarization systems mostly succeed at making the system more extractive rather than actually improving abstraction. Overall, chapter 2 makes important contributions to advancing faithfulness evaluation in summarization research.

While improved evaluation metrics and frameworks can lead to more accurate assessments of system performance, addressing the cause of faithfulness issues requires developing better attribution methods. In chapter 3, we focus on building methods to understand potential causes for observed faithfulness errors in summarization systems. Current summarization systems are often built by finetuning pretrained language models on summarization datasets. As such, errors in generated summaries could stem from several factors, including deficiencies in the training data, biases inherited from the pretraining process, and limitations of the finetuning methodology. In §3.1, we propose a new method for attributing faithfulness errors in generated output to the training data that improves upon the existing state of the art. Then, in §3.2, we propose an evaluation framework for measuring how distributional biases from the pretrained language model lead to hallucinations in summarization. We use this framework to assess multiple finetuning strategies and their effectiveness in reducing hallucinated content. Overall, chapter 3 makes important contributions towards explaining the origins of faithfulness issues in modern summarization systems.

In chapter 4, we build on the findings from the previous two chapters and develop techniques to mitigate faithfulness errors in summarization systems. §4.1 explores whether controlling abstractiveness can be a viable strategy for improving overall faithfulness. In particular, we propose a method to adaptively determine the appropriate level of abstractiveness for a given input, allowing the system to fall back to more extractive strategies when it's unable to faithfully abstract. We show that our approach outperforms prior work and can improve overall faithfulness without simply resorting to more extractiveness. In §4.2, we use our proposed error attribution method from §3.1 to trace faithfulness errors in generated outputs back to deficiencies in the training data. We show that training on a cleaned dataset leads to models with significantly reduced faithfulness errors in their generated outputs. Overall, Chapter chapter 4 makes important strides towards practically addressing faithfulness issues by leveraging insights around adaptive abstractiveness and data cleaning.

The success of pretrained language across natural language processing tasks has led to increased efforts to scale up language models in terms of parameters and dataset size. This has ushered a new era of Large Language Models (LLMs) (Brown et al. 2020; Chowdhery et al. 2022; Bommasani, Hudson, et al. 2021). Further tuning of these LLMs on a diverse set of tasks using both supervised finetuning and reinforcement learning (Sanh et al. 2021; Y. Wang et al. 2022; Ouyang et al. 2022; Yuntao Bai et al. 2022) has produced models capable of strong performance on many tasks without additional supervision. In chapter 5, we benchmark leading LLMs on few-shot and zero-shot summarization to assess their ability to summarize documents faithfully. §5.1 shows that state-of-the-art LLMs can produce faithful summaries of news articles on par with human-written ones. However, §5.2 finds that the strong performance in news summarization does not translate to the summarization of narratives. While LLMs can generate fluent and coherent summaries of narratives, they tend to contain many faithfulness errors. Moreover, the fact that generated summaries appear superficially sound further complicates evaluation – we show that crowd-workers cannot reliably identify faithfulness errors in the generated summaries. Overall, chapter 5 provides key insights into the capabilities and limitations of LLMs for summarization across domains, highlighting lingering challenges in faithfulness.

1.2 Contributions

This thesis makes the following key contributions:

- We demonstrate that existing faithfulness evaluation metrics are unreliable and suffer from spurious correlations. We then propose a new metric that addresses the spuriousness issue.
- We find that prior work for improving faithfulness primarily relies on increased copying and, therefore, may not be contributing to the goal of improving abstraction. We then propose a new framework to measure faithfulness that takes into account extractiveness.
- We develop a novel method for error attribution that allows us to trace faithfulness errors back to noisy training examples and show that it significantly outperforms prior methods for error attribution.
- We show how biases from pretrained models can propagate down to the summarization task in the form of faithfulness errors (hallucinations).
- We propose a novel approach for adaptively selecting the appropriate level of abstractiveness for a given input article, which leads to a more faithful and abstractive summarizer.
- Using our error attribution method, we clean existing datasets by tracing faithfulness errors back to problematic training instances. We show that models trained on the cleaned datasets are significantly more faithful.
- We benchmark LLMs for summarization and show that while they reliably summarize news articles, they struggle to faithfully summarize narratives.

Chapter 2: Evaluation of Faithfulness

Reliable, automated evaluation metrics are crucial for the rapid development of better summarization systems. Traditionally, the field has opted to evaluate summarization systems using reference-based evaluation – i.e., measuring similarity between system-generated summaries and human-written, gold summaries – with ROUGE (Lin 2004) being the metric of choice. ROUGE assigns a similarity score between generated output and reference(s) by computing n-gram overlap between the sequences. Simple n-gram matching, however, is not a suitable measure for measuring semantic overlap between abstractive summaries and references. Recent work attempts to address this shortcoming by relying on semantic representations from pretrained language models to compute the similarity between generated text and references, reporting significantly higher correlations with human judgment than simple overlap-based measures like ROUGE (T. Zhang, Kishore, et al. 2019; Hara et al. 2019).

Even with the improved reference-based evaluation metrics proposed by recent work, there are fundamental challenges with relying on referenced-based evaluation to measure the faithfulness of summaries. First, reference-based evaluation is an indirect measure of the faithfulness of a summary. If the reference-based evaluation score of a summary is high and the reference is faithful, then the generated summary is likely faithful. This assumption can be particularly problematic for current datasets as they are noisy, and the references may not always be faithful to the article (Maynez et al. 2020; Kang and T. B. Hashimoto 2020). Second, reference-based measures of faithfulness are confounded with content selection. A low reference-based score does not necessarily mean that the generated summary is unfaithful; it's possible that the generated summary simply focused on different details than the reference summary. Finally, collecting human-written references can be costly, especially since we would need to collect a large number of references in order to account for the content selection confound. Given these issues, it is unsurprising that recent work has shown that referenced-based evaluation metrics correlate poorly with human judgment of faithfulness (Durmus, H. He, et al. 2020; A. Wang et al. 2020; Maynez et al. 2020).

In §2.1, we first look at a proposed solution to the above challenges, namely learned reference-free evaluation metrics, and explore whether these metrics are actually able to capture the semantics associated with faithful summaries. Through our analyses, we show that the effectiveness of existing reference-free metrics may be overstated and that they may primarily rely on spurious correlations with extractiveness in evaluation datasets. When the evaluation distribution does not correlate with extractiveness, as is the case when evaluating systems that are close to the state-of-the-art, recently proposed reference-free metrics completely break down. To address this issue, we design a novel metric that is less reliant on spurious measures in the dataset and show that it can improve upon the existing state-of-the-art metrics.

The strong correlation between faithfulness and extractiveness not only makes it challenging to design good evaluation metrics but also makes it difficult to compare systems that operate at different levels of extractiveness. Current practice in the field is to report absolute faithfulness scores (along with ROUGE scores) for proposed systems; models with higher absolute faithfulness scores are considered better systems overall. In §2.2, we show that simply comparing two systems based on faithfulness scores paints an incomplete picture and may be detrimental to the goal of building better abstractive summarization systems. We propose a novel framework for evaluating systems that takes into account the extractiveness of generated summaries. Using this framework, we show that recently proposed methods for improving the faithfulness of summarization systems are mostly making the system more extractive rather than improving abstraction.

2.1 Reference-free Evaluation of Faithfulness

This is based on Durmus, Ladhak, et al. (2022), which was work done in collaboration with researchers from Stanford University. I was an equal contribution first author; ordering was decided via coin flip.

Given the issues highlighted above, recent work has turned to reference-free evaluation in order to measure faithfulness (Durmus, H. He, et al. 2020; Kryscinski, McCann, et al. 2020). Two main categories of reference-free faithfulness evaluation metrics have been proposed by the prior work: question-answering (QA) based and entailment-based metrics. QA-based metrics propose to extract questions from the generated summary and answer these questions both with the generated summary and the source article (Durmus, H. He, et al. 2020; A. Wang et al. 2020; Nan, Nogueira dos Santos, et al. 2021). They then compute the consistency of the answers generated using the generated summary and the source article to assess the summary's faithfulness. Entailment-based metrics aim to evaluate whether the source article entails a generated summary, and they are shown to be quite effective in identifying faithfulness issues (Pagnoni et al. 2021). Given their effectiveness, our work focuses on entailment-based metrics.

Many of these reference-free evaluations achieve remarkably high correlations with human evaluations, raising hopes that they may soon become a viable alternative to expensive human evaluations (Kryscinski, McCann, et al. 2020; Goyal and Greg Durrett 2020; Sinha et al. 2020; Phy et al. 2020; Gao et al. 2020). However, simply looking at the correlation with human scores may not be sufficient to determine the efficacy and robustness of an evaluation metric. In our work, we study recently proposed reference-free evaluation metrics of text summarization and dialog generation. We find that it is possible to achieve similar levels of correlation with human judgment using simple spurious correlates such as word overlap, length, and perplexity. Furthermore, we find that the learned metrics have a relatively high correlation with spurious correlates compared to human scores, suggesting that these metrics may rely heavily on spurious correlations. This may be a potential explanation for the robustness issues (e.g., inconsistent performance across evaluation sets) observed in recent work despite the seemingly high reported correlations with human judgments (Gabriel et al. 2021; Y.-T. Yeh et al. 2021).

We further analyze reference-free faithfulness evaluation metrics and show that reliance on spurious correlations leads to errors in model selection and development. First, we show that word overlap, a spurious correlate for the task, does as well as recently proposed reference-free metrics at system-level ranking. Then, we look at rankings amongst systems that are relatively abstractive and faithful, i.e., the current state of the art, and find that these learned metrics perform significantly worse for these systems. This is because word overlap is not a good measure for ranking these systems in terms of their faithfulness since all of these systems have similarly low word overlap. This suggests that we need metrics that are not overly reliant on word overlap in their faithfulness prediction.

2.1.1 Example-level and System-level Evaluation of Faithfulness Metrics

We begin by defining the task of reference-free evaluation, as well as the *example-level* and *systems-level* evaluation of these metrics.

We define a reference-free evaluation metric as a function $F(x, y)$ that can assign a quality score to an output sequence y for a given input sequence x. The goal of a reference-free evaluation metric $F(x, y)$ is to assign high scores to desirable outputs y for some attribute, such as the faithfulness of a summary. Measuring the quality of this metric is challenging, and prior work has relied upon correlation to human judgments $H(x, y)$.

Example-level evaluation: A number of existing reference-free evaluations rely upon a procedure which we call *example-level* human correlations (Alexander R Fabbri et al. 2020; Phy et al. 2020; Sinha et al. 2020), which measures the effectiveness of a metric by computing a Pearson or Spearman correlation $\text{corr}_{p_{\text{eval}}}(H(x, y), F(x, y))$ over some sampled evaluation data $p_{\text{eval}}(x, y)$.

System-level evaluation: An alternative approach to evaluation is *systems-level* rankings (Mathur et al. 2020; Kocmi et al. 2021), which we define as the ability to identify which model is better amongst a set of models M. F is evaluated via its accuracy in matching human evaluation H on all pairs $(m_i, m_j) \in M \times M$ where $m_i \neq m_j$.

The definitions of example and system-level correlations suggest that evaluations of these metrics may have a strong dependence on the example and systems distributions $p_{\text{eval}}(x, y)$ and M. As an example, consider an evaluation of dialogue response quality. Building a truly accurate predictor for dialogue response quality is challenging, but if $p_{\text{eval}}(x, y)$ consists of all either professionally written examples or ungrammatical nonsense, a simple grammar checker would perform exceedingly well.

This is an instance of what is called a spurious correlation. More formally, we define this as some attribute $S(x, y)$ which is correlated with H in $p_{\text{eval}}(x, y)$ but is not correlated with H for a carefully constructed test distribution $p_{\text{test}}(x, y)$. We say that F is *spuriously correlated* with S if:

1. F and H are highly correlated under $p_{\text{eval}}(x, y)$ but not under $p_{\text{test}}(x, y)$.

2. F remains correlated with S under $p_{\text{test}}(x, y)$.

2.1.2 Example-level Analysis

In this section, we look at example-level Spearman correlations with human judgments for reference-free evaluation metrics that have been proposed for summarization and dialog generation. We compare the metrics to spurious correlates, such as word overlap, length, and perplexity, in order to understand whether the metrics can perform better than these simple measures. We also measure to what extent the proposed metrics are correlated with these spurious measures.

2.1.2.1 Faithfulness Evaluation in Text Summarization

State-of-the-art text summarization models are capable of producing fluent summaries. However, they suffer from generating information that is not consistent (i.e., unfaithful) with the information in the source article (Z. Cao et al. 2018). Prior work showed that reference-based metrics are not able to capture such consistency errors (Falke et al. 2019). This motivated researchers to build evaluation metrics to capture these faithfulness issues since collecting human evaluations for faithfulness is expensive and time-consuming (A. Wang et al. 2020; Durmus, H. He, et al. 2020; Kryscinski, McCann, et al. 2020; Goyal and Greg Durrett 2020).

In this section, we analyze recently proposed reference-free faithfulness evaluation metrics and compare their performance against the spurious correlate of word overlap. Furthermore, we analyze the correlation between the learned metrics and word overlap to understand to what extent these metrics rely on spurious correlations. We focus on learned entailment-based faithfulness evaluation metrics due to their high performance in identifying faithfulness issues (Pagnoni et al. 2021). In particular, we evaluate FactCC (Kryscinski, McCann, et al. 2020) and DAE (Goyal and Greg Durrett

2021), which have been shown to achieve higher example-level correlations with human judgments than existing faithfulness evaluation metrics (Pagnoni et al. 2021).

FactCC. Kryscinski, McCann, et al. (2020) proposed an entailment-based method where they train a BERT-based model to predict whether or not the source article entails a summary. To train this model, they generate synthetic training data by applying a set of transformations to source article sentences in order to get article, summary pairs. They evaluate their approach on the CNN/DM dataset (See et al. 2017) and report a high accuracy on example-level comparisons on a human-annotated test set.

DAE. Goyal and Greg Durrett (2021) collected human annotations at the word-level and arc-level to study faithfulness at a finer granularity. They also trained a dependency arc entailment model for faithfulness detection (Goyal and Greg Durrett 2020). They evaluate on the same test set as Kryscinski, McCann, et al. (2020) and report improved results over FactCC.

We look at how these learned, reference-free metrics compare with word overlap – a simple spurious correlate. One simple measure of whether a generated summary is faithful is to look at its word overlap with the source article; summaries with a higher word overlap are more likely to be faithful (Ladhak, Durmus, H. He, et al. 2022). However, this measure of faithfulness is spurious because it cannot distinguish between faithful and unfaithful summaries that have similar word overlap. In particular, we look at two metrics of word overlap following Grusky et al. (2018): *coverage* and *density*. *Coverage* measures the percentage of the words in the summary that are also present in the article. *Density* instead looks at the average length of the segments in the summary that are extracted from the article.

Metric	Human	Density
FactCC	0.36	**0.59**
DAE	0.38	**0.76**

Table 2.1: Correlation of FactCC and DAE scores with humans vs density. Both learned metrics have a significantly higher correlation with density than human scores.

Results. We use the large-scale faithfulness human annotations collected by Alexander R Fabbri et al. (2020) for 16 summarization models on the CNN/DM dataset (See et al. 2017) for

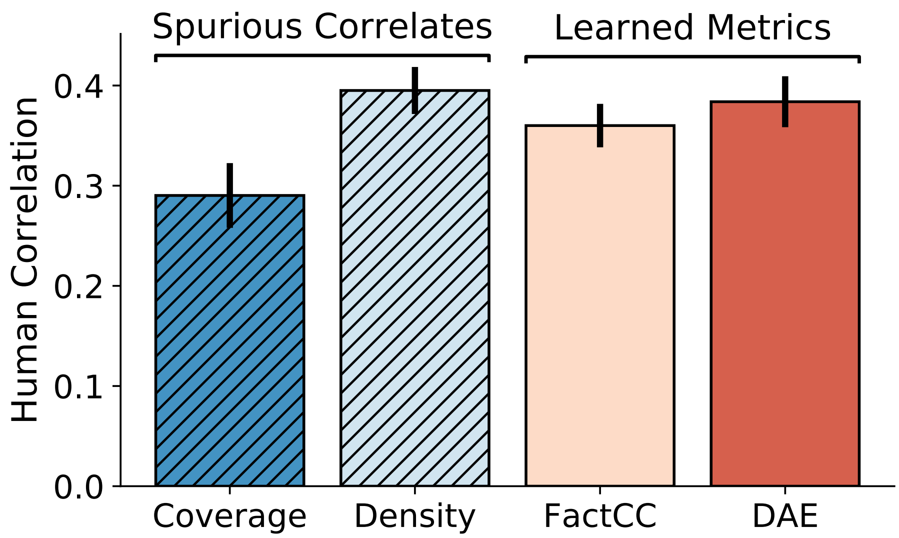


Figure 2.1: Correlation of the spurious correlates and learned metrics with human scores. Density, a spurious correlate, performs similarly to DAE and performs significantly better than FactCC.

our analysis. Figure 2.1 shows the example-level correlations with human scores for each of the factuality metrics as well as the spurious correlates. We note that *density* has a similar correlation with human scores as DAE, and is significantly[1] better than FactCC. This result is alarming because *density* is a spurious correlate, yet it can achieve similar performance as the metrics that have been trained for faithfulness evaluation.

Moreover, we also see that both FactCC and DAE have a significantly higher correlation with *density* than they do with human scores (Table 2.1). This indicates that these metrics may rely upon spurious correlations and are not yet capturing a deeper understanding of faithfulness.

2.1.2.2 Learned Metrics for Dialog Generation

To show that the spuriousness issue exists for learned reference-free evaluation metrics in general, beyond just summarization, we also evaluate recently proposed reference-free metrics. Dialog generation systems need to be able to generate a response given the dialog context. The ability to automatically evaluate the quality of a response is essential for building dialogue systems.

[1] All numbers reported in the paper are bootstrap means over 1000 bootstrap samples. We use a one-tailed percentile bootstrap test to determine significance at $\alpha = 0.05$.

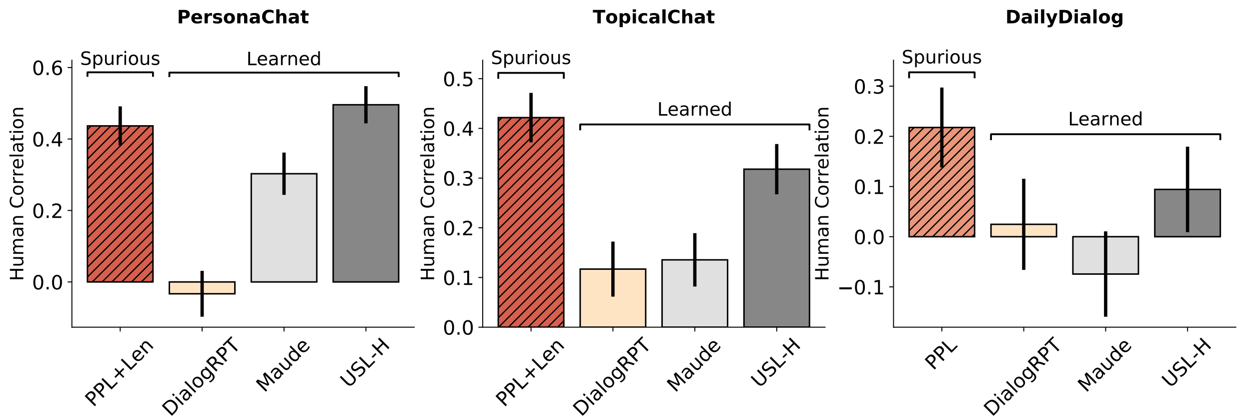


Figure 2.2: Correlation of the spurious correlates and learned metrics with human scores. PPL+Len represents a simple combination of perplexity (PPL) and length features. The best spurious correlate performs significantly better than all learned metrics on TopicalChat and performs similarly to the best-learned metric on PersonaChat and DailyDialog.

C.-W. Liu et al. (2016) show that referenced-based evaluation metrics do not correlate well with human judgments of response quality. This has led to an increased interest in reference-free evaluation metrics for evaluating dialogue response quality.

Similar to our analysis in section 2.1.2.1, we aim to look at recently proposed metrics for reference-free evaluation, along with spurious correlates for dialog response quality, and compare them against human judgments.

DialogRPT. Gao et al. (2020) finetune GPT-2 to predict the different types of human feedback (replies, upvotes, etc.) in Reddit threads and combine these to form a composite score for response quality. They evaluate their approach on the Reddit data that they collected and show that their method achieves higher example-level agreement with human judgments than baseline metrics.

MAUDE. Sinha et al. (2020) propose a model that encodes each utterance in the dialog context using a pretrained BERT model and leverages the temporal transitions between them to score a response. They add noise to existing dialog responses to create negative examples and train their system to distinguish them from valid responses using noise contrastive estimation (NCE). They evaluate their model on the PersonaChat (Saizheng Zhang et al. 2018) dataset and report improved example-level Spearman correlation with human judgments compared to baseline metrics.

USL-H. Phy et al. (2020) decompose response quality into three aspects and train a model to score a response along each of these aspects. They then combine the scores hierarchically into one composite score for response quality. They evaluate their metric on the DailyDialog (Y. Li et al. 2017) dataset and report significantly higher example-level correlations than previous baseline metrics.

MNLI+Adv. Dziri et al. (2021) introduce an entailment-based metric that evaluates the groundedness of a dialog response, i.e., whether the generated response is consistent with the information in the provided external context, such as a Wikipedia article. They trained their metric on automatically generated adversarial data by applying perturbations to the evidence. They further collect human annotations for the various aspects of dialog generation, such as entailment, genericness, etc., and show that their method is more effective in accurately categorizing the generations than existing entailment models.

To assess these metrics, we look at two spurious correlates for dialog quality – perplexity and length of the generated output – as well as a simple combination of two measures. We compute perplexity using a pretrained GPT-2 language model (Radford et al. 2019). Perplexity (PPL) and length are spurious correlates since they do not account for the dialog context, and therefore it is possible to have high-quality and low-quality responses with similar perplexities/lengths. For groundedness evaluation, we look at the same word overlap measures as we did for summarization, i.e., *density* and *coverage*, and we measure overlap between the response and the provided external evidence.

Results. We evaluate metrics[2] for response quality estimation on three popular multi-turn dialog datasets – DailyDialog, which contains dialogs about everyday topics (Y. Li et al. 2017), TopicalChat, which contains dialogs conditioned on a set of 8 broad topics (Gopalakrishnan et al. 2019), and PersonaChat, which contains dialogs conditioned on personas (Saizheng Zhang et al. 2018).

To evaluate the recently proposed metric for response groundedness, we use human annotations collected by Dziri et al. (2021) on Wizard of Wikipedia (Dinan et al. 2019), a dataset that consists of

[2]We use the code provided by Y.-T. Yeh et al. (2021) for these experiments.

dialogues conditioned on information from Wikipedia articles. In particular, we use their entailment annotations, where human annotators judge whether or not the external evidence entails a generated response.

		Human	Perplexity	Length	PPL+Len
PersonaChat	DialogRPT	-0.033	-0.017	**0.086**	0.068
	Maude	0.303	**0.373**	-0.089	0.137
	USL-H	0.496	0.092	**0.506**	0.469
TopicalChat	DialogRPT	0.117	-0.011	0.272	**0.276**
	Maude	0.135	**0.243**	-0.191	-0.148
	USL-H	0.318	0.037	**0.359**	0.355
DailyDialog	DialogRPT	0.025	-0.182	**0.359**	0.270
	Maude	-0.074	-0.076	**0.102**	0.033
	USL-H	0.094	0.048	-0.208	**-0.236**

Table 2.2: Correlation of the metrics with human scores and spurious correlates. Reference-free evaluation metrics have a higher correlation with spurious correlates than human scores.

Figure 2.2 shows the correlations with the human scores and the spurious correlates for the dialog generation evaluation metrics. In DialyDialog, we find that perplexity achieves a similar correlation with human judgments as USL-H. In TopicalChat, perplexity or length alone does not beat out any of the learned metrics; however, combining the two measures achieves a significantly better correlation with humans than learned metrics. In PersonaChat, USL-H achieves the highest correlation with human judgment, though the combined PPL+Len score is close. We observe that USL-H is more consistent than the other reference-free metrics and achieves significantly higher correlations with human scores than MAUDE and DialogRPT for PersonaChat and TopicalChat. We further find that the reference-free metrics have a higher correlation with the spurious correlates than the human scores (Table 2.2), which again suggests that these learned metrics may be relying upon spurious correlations.

For groundedness evaluation[3], both *coverage* and *density* achieve significantly higher correlation with human scores than MNLI+Ad and USL-H. Furthermore, MNLI+Ad and USL-H get a higher correlation with these spurious correlates than human scores (Figure 2.3).

[3]We do not include MAUDE and DialogRPT results for this task since they perform significantly worse.

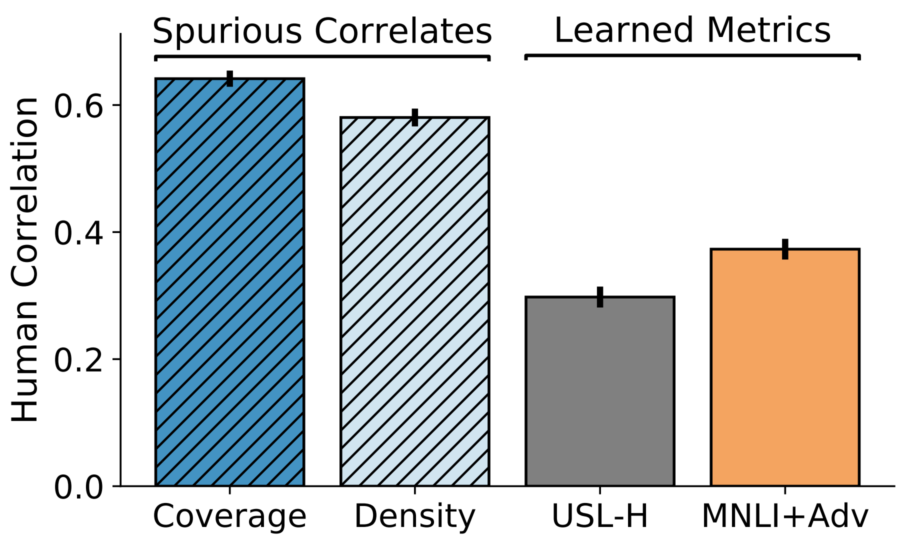


Figure 2.3: Correlation of the spurious correlates and learned metrics with human scores on groundedness evaluation. Both coverage and density get significantly higher correlations with human scores than the learned metrics.

Despite relatively high correlations on their original datasets, these metrics perform similarly to simple spurious correlations on other datasets. In order to better understand the effectiveness of these reference-free evaluation metrics, we suggest that future research includes comparisons to potential spurious correlates and that research communities come up with a set of potential standard spurious correlates.

Metric	Human	Coverage	Density
USL-H	0.298	0.467	**0.515**
MNLI+Adv	0.373	0.451	**0.514**

Table 2.3: Correlation of USL-H and MNLI+Adv scores with humans vs. coverage and density. Both learned metrics have a significantly higher correlation with density than human scores.

2.1.3 System-level Analysis

Our example-level analysis demonstrates that recently proposed learned evaluation metrics achieve worse correlations with human scores than spurious correlates for almost all the settings. Since an important goal of building these metrics is to be able to rank arbitrary systems, we analyze

Model Name	Paper
M0	Lead-3 baseline
M1	Q. Zhou et al. (2018)
M2	Y. Dong, Shen, et al. (2018)
M5	Y. Wu and Hu (2018)
M8	See et al. (2017)
M9	Y.-C. Chen and Bansal (2018)
M10	Gehrmann, Deng, et al. (2018)
M11	Kryściński, Paulus, et al. (2018)
M12	Hsu et al. (2018)
M13	Pasunuru and Bansal (2018)
M14	H. Guo et al. (2018)
M15	Jiang and Bansal (2018)
M17	Raffel et al. (2019)
M20	Ziegler et al. (2019)
M22	Lewis et al. (2020)
M23	J. Zhang et al. (2020)

Table 2.4: Systems included in the evaluation set by collected by Alexander R Fabbri et al. (2020).

whether these concerns we observe at the example level manifest into harms at the system level (i.e., ranking systems incorrectly). In order to study this, we need a large collection of human evaluation data across a wide range of systems. The dataset collected by Alexander R Fabbri et al. (2020) contains human evaluations for faithfulness across 16 summarization systems, making it suitable for this study. The set of systems that are in the evaluation dataset is shown in Table 2.4. This dataset allows us to evaluate faithfulness evaluation metrics further in order to assess how accurately they rank summarization systems in terms of faithfulness.

We first measure pairwise ranking accuracy across all the systems in the dataset. We find that system-level rankings suffer from a similar issue as the example level correlations: density and coverage appear as spurious correlations (Table 2.5). From this observation, we perform a finer-grained analysis and show that these factuality metrics fail on the most important subset of model comparisons: abstractive but faithful summarization system (AF) – where the current state-of-the-art abstractive summarization systems fall. This is the set of systems shown in the blue box in Figure 2.4.

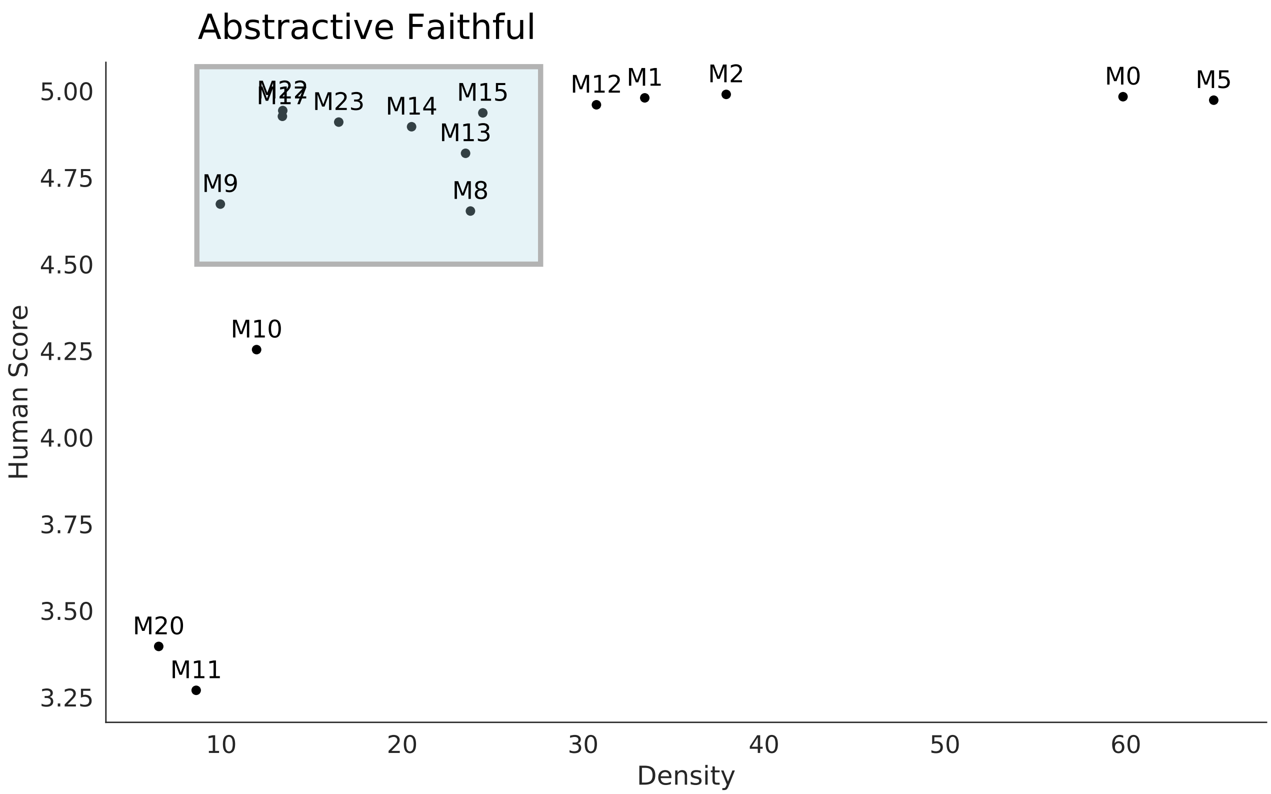


Figure 2.4: Density and human scores for summarization systems. We analyze the accuracy of the metrics in ranking all the systems vs. ranking the systems within the abstractive faithful group, shown in the blue box. Abstractive faithful systems have a faithfulness score higher than 4.5 (out of 5) and a density lower than 30.

2.1.3.1 Results

Both faithfulness metrics perform relatively well when we look at pairwise ranking accuracy across all pairs of models (Table 2.5). However, they are unable to improve over *density*, which achieves the highest overall accuracy. When we look at ranking within the AF group, we see *density* is no longer a good measure for the faithfulness of a system since these systems are relatively close in terms of density. Similarly, the performance of the learned metrics drops significantly, which is an expected result since our analysis in section 2.1.2.1 showed that both FactCC and DAE are spuriously correlated with density. We claim that our system-level analysis is further evidence that these metrics may rely heavily on simple spurious measures such as word overlap.

	All Pairs	Within AF
Coverage	56.54	26.60
Density	**81.01**	**40.45**
FactCC	78.87	38.26
DAE	80.39	37.88

Table 2.5: Accuracy of pairwise ranking across all the systems and within Abstractive Faithful (AF). We observe that the ranking accuracy of all metrics is significantly lower for systems within AF compared to all pairs. Density performs as well as the best learned metric (DAE) in both cases.

These results highlight the importance of performing analyses across different distributions of systems. If we were looking at just the overall ranking accuracy of the metrics, we would conclude that DAE and FactCC correctly measure faithfulness. However, on closer examination, we see that both metrics perform relatively poorly in ranking AF systems, which is arguably the most crucial group since most state-of-the-art systems operate in this regime, and there is substantial interest in building abstractive and faithful summarization systems.

2.1.4 Improved Reference-free Metric via Adversarial Training

In our earlier example-level analysis, we found that learned metrics have a higher correlation with spurious correlates than human judgment. We further saw in our system-level analysis that learned metrics for faithfulness are unable to outperform density. One natural question that follows is whether we can build metrics that do well at the systems level by learning representations that rely less on spurious correlates.

In order to do this, we train an entailment-based model using the synthetically generated data from FactCC in an adversarial setup similar to Ganin et al. (2016). In particular, our approach augments the standard faithfulness predictor with a density predictor that tries to predict the density of the summary from the model's internal representation. We use this density predictor as an adversary, and our goal is to predict faithfulness while ensuring that it is difficult to predict density using this exact representation. To achieve this, the gradients from the density predictor are reversed, which makes it harder to predict the density from the encoder's representation and thus makes the faithfulness predictions less reliant on density. The model architecture is shown in Figure 2.5. We

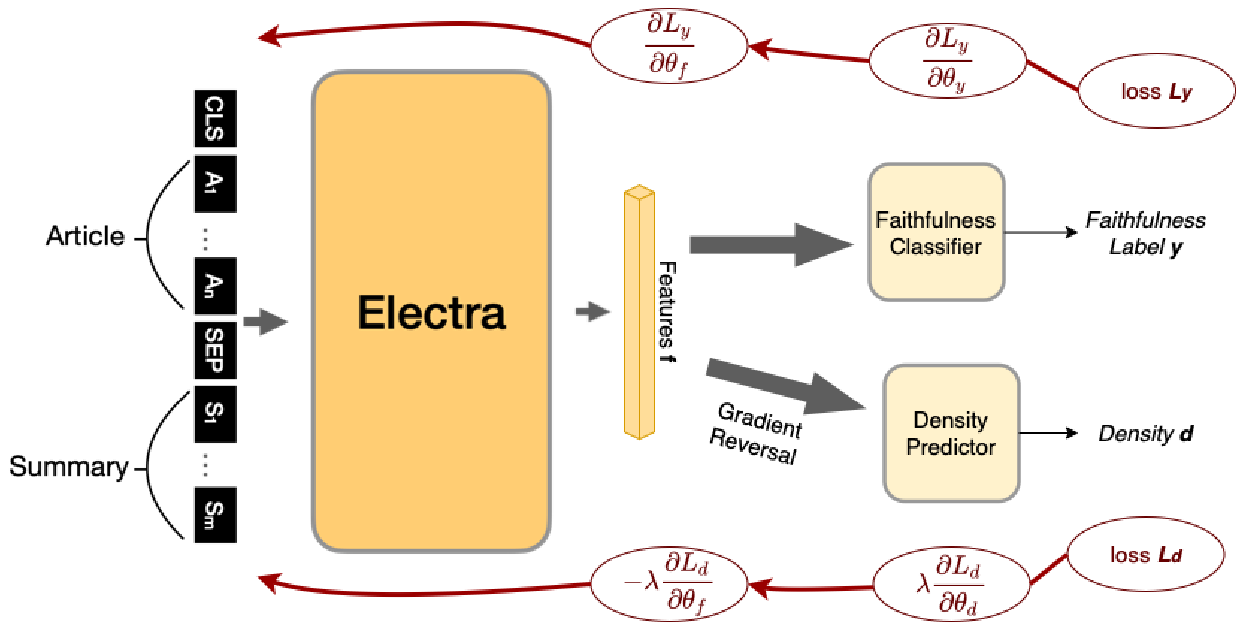


Figure 2.5: Architecture of adversarial model. The input sequence is first encoded via a pretrained Electra model, and the representation is used for both faithfulness classification and density prediction. Gradients from the density predictor are reversed in order to make updates to the encoder's parameters, forcing the model to learn representations that are not predictive of density.

initialize the parameter λ to 0 and gradually increase it to 1, following the schedule detailed in Ganin et al. (2016).

We finetune a pretrained Electra model (Clark et al. 2020) using the transformers library (Wolf et al. 2020) for this task. We chose Electra in order to match the model architecture in DAE. Since the original FactCC metric was finetuned on BERT, we also finetuned our own version of FactCC on Electra (FactCC-Electra) as an ablation. Our adversarially trained model is essentially the same as FactCC-Electra but with an additional adversarial head for predicting density.

	All Pairs	Within AF
FactCC-Electra	77.85	27.70
FactCC	78.87	38.26
DAE	80.39	37.88
Adversarial	**85.27**	**59.20**

Table 2.6: Pairwise ranking accuracy for systems across All Pairs vs. Within Abstractive Faithful (AF) for DAE and Adversarial. Adversarially trained metric performs significantly better for the systems within AF than previously proposed metrics.

2.1.4.1 Resutls

We note that the FactCC-Electra model performs worse than the original FactCC, which is consistent with the findings in Goyal and Greg Durrett (2021). Our adversarially trained metric has a significantly lower example-level correlation with density (27.71%), as compared to FactCC (59.10%) and DAE (76.37%). We find that the adversarial model[4] can achieve a significantly better performance than existing learned evaluation metrics in ranking systems within the abstractive faithful (AF) group (Table 2.6). This suggests that it is possible to learn effective metrics that are not overly reliant on spurious correlates. Furthermore, our metric is also effective in the overall pairwise ranking of the systems achieving 85.27% accuracy.

2.2 Faithfulness-Abstractiveness Trade-off

This is based on Ladhak, Durmus, H. He, et al. (2022), which was work done in collaboration with researchers from Cornell University and NYU.

In the previous section, we highlighted the importance of taking extractiveness into account when designing and evaluating faithfulness evaluation metrics. We further show how this can impact the evaluation of abstractive summarization systems. While recent work has proposed methods to improve the faithfulness of abstractive summarization systems (Kang and T. B. Hashimoto 2020; Goyal and Greg Durrett 2020; S. Chen et al. 2021), it is unclear whether the improvement comes from an increased level of extractiveness, as one straightforward way of improving faithfulness of generated summaries is to copy a larger amount of content from the source article (i.e., more extraction). Thus, any methods that increase extractiveness, intentionally or not, would improve faithfulness. In our work, we argue that in order to make progress in abstractive summarization, it is important to tease apart faithfulness improvements due to increased extractiveness versus improvements due to improved abstraction.

In order to tease this apart, we develop a framework for evaluating progress in faithfulness by considering the *effective faithfulness*, i.e., the improvement in faithfulness over a baseline system

[4] Our adversarially trained model can be found at https://github.com/esdurmus/adversarial_eval.

(*control*) operating at the same level of extractiveness. In particular, we split the training examples into groups according to their extractiveness of the summary and train *control models* for each group. Each of these models corresponds to a specific trade-off between abstractiveness and faithfulness, forming a *trade-off curve* indicating how much faithfulness can be improved solely by increasing extractiveness. Systems that improve *effective faithfulness* should lie above this curve. Using this framework, we show that the improved faithfulness of recently proposed methods comes mainly from an increased extractiveness.

2.2.1 Datasets

We conduct our study on two English abstractive summarization datasets, one from the news domain and one from a non-news domain. For the news domain dataset, we decided against using the popular CNN/Dailymail dataset since its reference summaries tend to be very extractive (Kedzie, K. McKeown, et al. 2018; Bommasani and Cardie 2020), making it a poor choice for studying faithfulness in abstractive summarization. Similarly, we also decided against using XSum, another popular news summarization dataset, since almost 77% of the gold reference summaries contain hallucinations (Maynez et al. 2020). Instead, we opted for Gigaword and Wikihow, which are datasets with substantial abstraction, with fewer hallucination problems in their references compared to XSum. Gigaword reference summaries have substantially fewer hallucinations than XSum (Kang and T. B. Hashimoto 2020), and WikiHow summaries tend to be of higher quality since they are written and curated by humans (Koupaee and W. Y. Wang 2018; Ladhak, Durmus, Cardie, et al. 2020).

Wikihow (Koupaee and W. Y. Wang 2018) is a dataset of how-to articles covering a diverse set of topics, collected from the wikihow.com website. Each article contains several paragraphs detailing step-by-step instructions for a procedural task. There are about 12M such paragraphs in the dataset, paired with a one-sentence summary.

Gigaword (Rush et al. 2015) is a headline generation dataset that contains around 4M examples extracted from news articles that were collected as part of the Gigaword corpus (Graff et al. 2003).

The model is tasked with generating the article's headline, given the first sentence.

Dataset Extractiveness. We follow the process detailed by Grusky et al. (2018), and use *extractive fragment coverage* and *extractive fragment density* as the measures of extractiveness of a given summary. Henceforth, we will refer to these as coverage and density, respectively. Coverage is the percentage of words in a summary that are from the source article. Density is the average length of the text spans copied from the document that are contained in the summary. A summary that copies larger chunks of text from the source article will have a higher density.

2.2.2 Analysis on Metrics of Faithfulness

Recent studies of faithfulness evaluation have proposed model-based automated metrics to detect whether a given summary is faithful to the source article. For example, Falke et al. (2019) **(Entailment)** have studied using pretrained entailment-based methods to assess the probability of the generated output being entailed by the source article. Kryscinski, McCann, et al. (2020) **(FactCC)** augment hallucinated summaries by applying rule-based transformations to the document sentences and train a BERT-based model to classify whether the generated output is faithful. Goyal and Greg Durrett (2021) **(DAE)** have collected fine-grained annotations to study word-, dependency- and sentence-level faithfulness and use these annotations to train a factuality detection model.

Figure 2.6 shows the relationship between the average coverage of the generated outputs (extractiveness) vs. average metric scores (faithfulness) assigned to various abstractive summarization models trained on Gigaword.[5] We observe that there is a positive correlation between extractiveness and faithfulness scores, as models whose generated summaries have a higher average coverage tend to also get higher scores for each of the faithfulness metrics. This correlation between exrtactiveness and faithfulness makes it unclear whether a model gets higher factuality scores simply because it is more extractive or it is capable of generating faithful summaries at the original level of extractiveness. This highlights the need to account for extractiveness to compare faithfulness across different abstractive summarization systems. Furthermore, given what we saw in §2.1, these

[5]These are the baseline and quartile models that are described in §2.2.3.1.

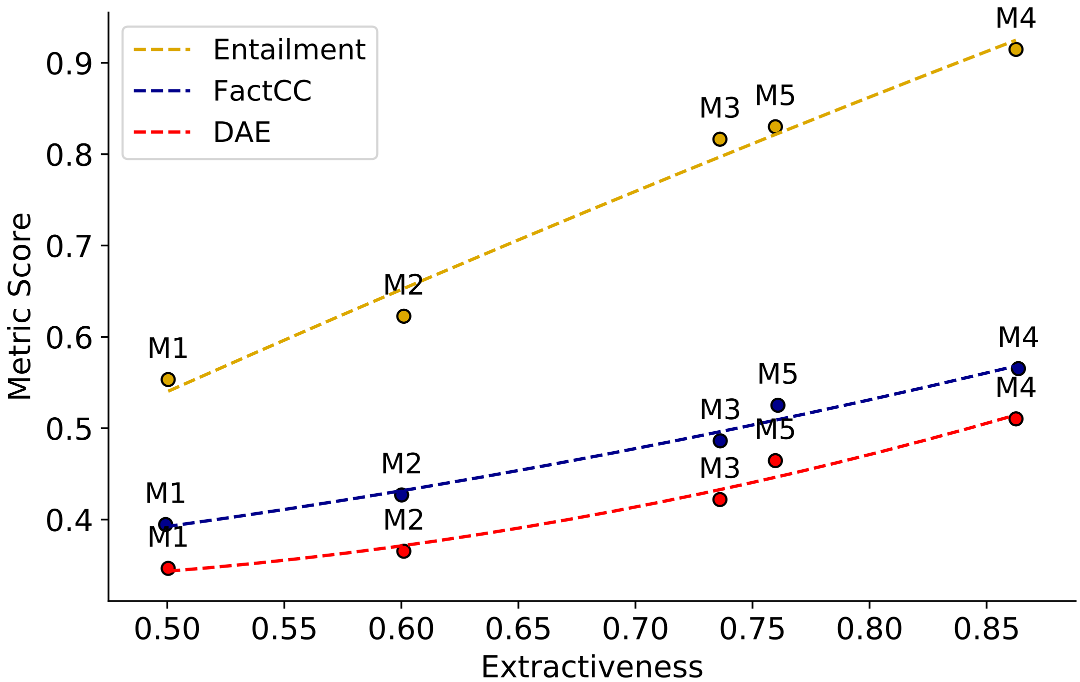


Figure 2.6: Extractiveness of generated outputs versus automated metric scores for Entailment, FactCC and DAE on the Gigaword dataset. We use *coverage* defined in Grusky et al. (2018) to measure extractiveness, where summaries with higher coverage are more extractive. We observe that automated metrics of faithfulness are positively correlated with extractiveness.

reference-free evaluation metrics rely mostly on spurious correlations and, therefore, may assign higher scores to systems that are more extractive on average, regardless of whether or not those systems are actually more faithful. As such, we will rely on human evaluations of faithfulness for the rest of the experiments in this section.

2.2.3 Effective Faithfulness

Given that extractiveness is confounded with faithfulness, we propose a framework for evaluating *effective faithfulness*, which takes into account the extractiveness of a system. In order to do this, we first need to determine the faithfulness of a system operating at a given level of extractiveness. We call this the *Faithfulness-Abstractiveness Tradeoff*, and we describe it further in §2.2.3.1. The *effective faithfulness* of a system is then simply the relative difference between the faithfulness score assigned to the system and the score of a system operating with the same average extractiveness according to the trade-off curve.

2.2.3.1 Faithfulness-Abstractiveness Tradeoff

In order to understand the effectiveness of a proposed system for improving faithfulness, we need to be able to account for its extractiveness. We finetune pretrained BART models (Lewis et al. 2020) for different levels of extractiveness, without any explicit recourse for improving faithfulness. We then use these systems to create a *faithfulness-abstractiveness trade-off curve* that can serve as a control to measure the *effective faithfulness* of summarization systems. Models that improve *effective faithfulness* should lie above the *faithfulness-abstractiveness trade-off curve*.[6]

Dataset	Quartile	# Examples	Article Length	Summary Length
Gigaword	Q1	985,931	30.58	8.03
	Q2	961,970	32.02	8.32
	Q3	952,833	31.77	8.41
	Q4	903,223	31.05	8.17
Wikihow	Q1	328,470	50.73	7.63
	Q2	221,452	75.69	7.40
	Q3	206,558	85.44	5.96
	Q4	243,837	92.09	5.49

Table 2.7: Data statistics for each quartile. *Length* corresponds to average # of words.

In particular, we sub-sample the training data into extractiveness quartiles by computing the coverage of the references with respect to the source articles. The number of examples, source article length, and target summary length for each quartile are shown in Table 2.7. To create the quartiles, we first compute the extractiveness (e_x) of the reference summary for each training example x, and compute the 25th (a), 50th (b), and 75th (c) percentile of the extractiveness of the

[6]Human evaluation data and trade-off curves can be found at https://github.com/fladhak/effective-faithfulness.

training data. The quartiles are then created as follows:

$$q1 = \{x \mid e_x \leq a\}$$

$$q2 = \{x \mid a < e_x \leq b\}$$

$$q3 = \{x \mid b < e_x \leq c\}$$

$$q4 = \{x \mid e_x > c\}$$

Note that it is possible for there to be several points at the boundary, and therefore there is an unequal number of examples in each quartile as shown in Table 2.7. For Gigaword, the article and summary lengths are very similar for each of the quartiles. For Wikihow, we observe that the article length is longer and the summary length is shorter for more extractive quartiles. We then finetune BART on each of these quartiles to obtain **quartile models** with varying levels of extractiveness. In addition, we also finetune BART on all of the data, which we call the **baseline**.

Article:

The Academy Selsey suffered extensive damage and much of its contents was destroyed in the blaze on 21 August. The school's 453 pupils are currently being taught at four locations in the Selsey area. As well as classrooms, the temporary buildings will house science labs, workshops and food technology areas, the school said. The buildings, some of which have already been placed on the site by two large cranes, had previously been used to house staff and students from Bohunt School in Worthing while building work was carried out. Tom Garfield, head teacher of the academy, said: "We are absolutely delighted to see the first temporary school buildings arrive on site. It's a great milestone for us, the staff and students alike." He said once the remaining structures had arrived over the next few days the school would begin preparing the rooms for teaching.

Output 1: The first temporary school buildings have arrived at the site of a school destroyed in a fire in Selsey.

○ Supported
○ Not Supported

Figure 2.7: An example from our human evaluation.

We follow a similar procedure as the prior work to collect human evaluations for the faithfulness of the generated summaries (Alexander R Fabbri et al. 2020). Given the source articles and generated summaries, we ask annotators to judge whether the generated summary is **supported** by the article. The output is supported by the article if all the information expressed by the output can also be inferred from the article. We ask annotators to ignore minor grammatical errors and focus on the information content of the generated summaries. Figure 2.7 shows an example from our human

Dataset	Model	Coverage	Faithfulness
Gigaword	Baseline	76.12	83.33
	Q1	50.25	71.83
	Q2	60.57	79.50
	Q3	73.64	86.67
	Q4	86.94	89.17
Wikihow	Baseline	88.28	82.52
	Q1	81.34	67.82
	Q2	85.34	76.21
	Q3	87.59	80.35
	Q4	90.19	91.08

Table 2.8: Coverage and faithfulness values of the baseline and each quartile model for Gigaword and Wikihow. Quartile models with higher coverage have higher faithfulness scores.

evaluation.

We evaluate 200 output summaries per system and each output is evaluated by three annotators. We restricted the study to the annotators with a high acceptance rate ($\geq 98\%$) and at least 5000 HITs to ensure annotation quality.[7] We follow prior work (Durmus, H. He, et al. 2020) and take the percentage of annotators who judge the summary as faithful to be the faithfulness score of a summary. To get the faithfulness score for a system, we average the summary scores across all 200 samples.

Table 2.8 shows the coverage and faithfulness scores for the baseline and the quartile models, where Q1 is the *most abstractive* and Q4 is the *most extractive* quartile. We observe that the models that are finetuned on more extractive quartiles produce outputs with significantly higher coverage and faithfulness scores. The baseline model generates relatively extractive outputs with coverage closest to Q3 on both Gigaword and Wikihow. Furthermore, we observe that the baseline model has a higher coverage than the model finetuned on Q3 but it has a lower faithfulness score for Gigaword.

Table 2.9 shows an article from the Wikihow dataset and corresponding output summaries generated by the baseline and each of the quartile models. We observe that the generated summaries are very similar in meaning; however, the output generated by the Q1 model includes a higher number of novel words (i.e. lower coverage) compared to the other models while staying faithful to

[7] We hired annotators from the USA, UK, and Australia. The data collection protocol was approved by IRB.

Article	Once you decide what to outsource, look for the right contractors. Start by asking for referrals from your own professional network. Talk to other business owners and professionals about how and where they outsource. You can also check professional associations or trade groups field in which you are trying to outsource work. Use other social media platforms such as Facebook or Twitter to advertise what you are looking for. Alternately, you can connect with contractors and freelancers on sites such as eLance, Guru and oDesk. These websites allow business owners to place an ad that describes what kind of work they need to have done, and contractors respond with their qualifications and rates. [TRUNCATED] ...
Baseline	Search for contractors and freelancers to outsource the work.
Q1	Conduct an initial search for qualified contractors and freelancers.
Q2	Search for qualified contractors and freelancers to work on your project.
Q3	Search for contractors and freelancers to do the work.
Q4	Look for contractors and freelancers to bid on the work.

Table 2.9: Example summaries generated by the baseline and quartile models for the article "*How to Outsource Small Business Tasks*" from the Wikihow dataset. The tokens that do not appear in the source article are indicated by green.

the article. Conversely, the Q4 model has a coverage of 1 in this example; all the words generated by this model are from the source article. On average, the Q1 model generates outputs that are more abstractive and less faithful while Q4 generates outputs that are more extractive and more faithful.

2.2.4 Evaluating Existing Methods for Effective Faithfulness

2.2.4.1 Baseline

For the **baseline**, we train a BART model on the entire training set, using the same hyperparameter settings from Lewis et al. (2020), without any mitigation for improving faithfulness. We compare this against two recently proposed methods for improving the faithfulness of summarization systems.

2.2.4.2 Loss Truncation

Kang and T. B. Hashimoto (2020) have proposed a method to adaptively remove high-loss examples to optimize the distinguishability of samples from the model and the reference. They have shown that the samples generated by this Loss Truncation model achieve higher factuality ratings compared to the baseline methods. We study this method to understand where it lies in terms of faithfulness-abstractiveness trade-off and whether it can achieve an improved *effective faithfulness* over the *control*.

2.2.4.3 Dependency Arc Entailment (DAE)

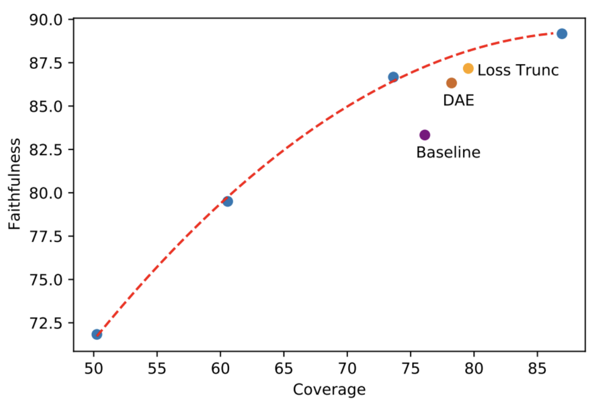
Goyal and Greg Durrett (2020) have proposed a factuality evaluation metric (DAE) that evaluates whether each dependency arc in the generated output is consistent with the input. They show that their proposed metric works better than existing factuality metrics, while also being able to localize the parts of the generated output that are non-factual. Goyal and Greg Durrett (2021) take advantage of DAE's ability to localize factuality errors and train a summarization model only on the subset of tokens that is deemed factual according to the DAE metric. We follow their methodology to train summarization models and assess them using our evaluation framework.

2.2.4.4 Results

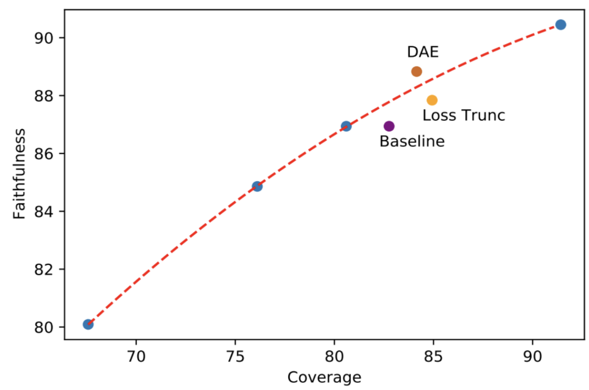
Figure 2.8 shows the abstractiveness-faithfulness trade-off curves (red dashed lines) for Gigaword and WikiHow that are generated using the quartile models. These curves represent the expected faithfulness for models operating at a given level of extractiveness for the given dataset. Proposed summarization systems can be measured against this curve by computing their average coverage (extractiveness) and average faithfulness scores on the evaluation set. We plot the **baseline** system, along with **Loss Trunc** and **DAE**, in order to see how these models stack up against the trade-off curve. If a model lies above the curve, it is improving *effective faithfulness*. If the model is below this curve, then it cannot improve the *effective faithfulness* and has a worse trade-off than the *control* operating at the same level of extractiveness. This means we can get a better model by training on a sub-sampled training set at the same extractiveness level.

We see in Figure 2.8 that the **baseline** is consistently sub-optimal according to our *effective faithfulness* measure. Furthermore, we see that while the recently proposed methods for improving faithfulness are better than the **baseline**, they are unable to consistently improve *effective faithfulness*. We can also see that both methods are also more extractive, on average, compared to the baseline.

This suggests that the improved faithfulness scores for these approaches, over the baseline, might be coming simply from increased extractiveness. They are, however, sub-optimal compared to the *control* operating at the same level of extractiveness. This shows that more work is needed to build systems that are able to improve *effective faithfulness*, and simply comparing absolute faithfulness scores is not the best way to compare systems.



(a) Trade-off and existing methods on **Gigaword**.



(b) Trade-off and existing methods on **Wikihow**.

Figure 2.8: Performance of existing approaches compared to the trade-off curve. The baseline as well as the recently proposed methods for improving faithfulness are below the curve and do not consistently improve effective faithfulness.

2.3 Related Work

There has been a lot of recent work in abstractive summarization showing that state-of-the-art systems suffer from generating inconsistent information with respect to the source article, despite their improved success in producing fluent summaries (Falke et al. 2019; Lux et al. 2020; Wilber et al. 2021). Since word-overlap based metrics such as ROUGE have low correlation with human scores of faithfulness (Kryscinski, Keskar, et al. 2019; Alexander R Fabbri et al. 2020), there has been significant effort to develop automated metrics that can detect such errors (C. Zhou et al. 2021; Gabriel et al. 2021; Pagnoni et al. 2021). For example, Falke et al. (2019), Maynez et al. (2020) and Goyal and Greg Durrett (2020) have proposed to assess faithfulness using entailment models, where a faithful summary should be assigned a high entailment score with respect to the original article. Kryscinski, McCann, et al. (2020) presented FactCC, a weakly-supervised BERT-based entailment model, by augmenting the dataset with artificial faithfulness errors. Durmus, H. He, et al. (2020) and A. Wang et al. (2020) proposed question-answering-based evaluation frameworks by automatically generating questions from the generated summary, and comparing the corresponding answers from both the source and the generated summary in order assess information consistency. Furthermore, several benchmarks have been proposed to evaluate the strengths and weaknesses of these evaluation metrics (Gabriel et al. 2021; Pagnoni et al. 2021).

The work in this chapter is related to these efforts of improving the evaluation of faithfulness in summarization systems but differs in two key ways. First, while the focus of most work on summarization evaluation metrics has been to propose new, improved metrics, the work in this chapter instead focuses on assessing the evaluation methodology of evaluation metrics in order to understand whether or not claims of improved metrics actually hold true. Second, while recent work on faithfulness evaluation simply compares systems using absolute faithfulness scores, we propose a more holistic evaluation framework that takes abstractiveness into account.

With respect to assessing the evaluation methodology of evaluation metrics, most existing work has focused on reference-based evaluation. For example, Mathur et al. (2020) takes a critical look at the use of example-level correlations to measure reference-based evaluation metrics in Machine Translation. They show that evaluating these metrics using example-level correlations can be sensitive to the presence of outliers which can lead to false conclusions about a metric's efficacy. Furthermore, Kocmi et al. (2021) show that proper assessment of evaluation metrics is crucial as uninformed use of automated metrics such as BLEU can lead to bad deployment decisions. Caglayan et al. (2020) has shown that automated reference-based evaluation metrics have robustness issues which can cause them to score generated outputs higher than human written outputs. Furthermore, Bhandari et al. (2020) has studied the limitations of reference-based evaluation metrics of text summarization, comparing these metrics across different datasets and application scenarios. In contrast, our work focuses on analyzing learned, reference-free evaluation metrics in summarization and dialog generation, accounting for potential spurious correlates for these evaluation tasks.

There has been some recent work comparing existing reference-free evaluation metrics for text summarization and dialog generation. Pagnoni et al. (2021) has measured the efficacy of existing reference-free faithfulness evaluation metrics of summarization on two different summarization datasets relying on example-level correlations. Similarly, Gehrmann, Adewumi, et al. (2021) has evaluated automated metrics of text summarization across a wide range of datasets. Gabriel et al. (2021) has proposed a meta-evaluation framework to evaluate the evaluation metrics looking at certain aspects of these metrics such as robustness, sensitivity, high correlation with human scores, etc., and measure existing evaluation metrics across these aspects. Y.-T. Yeh et al. (2021) perform a comprehensive study of existing dialog generation metrics across several different datasets and find that the performance of metrics varies widely across datasets.

Gabriel et al. (2021) and Y.-T. Yeh et al. (2021) are the most related to our work since they study the robustness of these metrics looking at their performance across different datasets. In our work, however, we explicitly study spurious correlations and show that these may potentially be contributing to the robustness issues. We further present initial promising results suggesting that controlling for these spurious correlates may result in more robust evaluation metrics.

Improving faithfulness of summarization systems is essential for deploying these systems in real-world scenarios, as such recent work has studied methods to improve the faithfulness of abstractive summarization systems (Matsumaru et al. 2020; Z. Zhao et al. 2020; Y. Dong, S. Wang, et al. 2020; Goyal and Greg Durrett 2021; Xu et al. 2020; S. Chen et al. 2021; Zhu et al. 2021). For example, Goyal and Greg Durrett (2021) train summarization systems by modifying the training objective to maximize the likelihood of the subset of summary tokens that are considered faithful according to their factuality detection model. Z. Zhao et al. (2020) specifically target hallucination of quantities in generated summaries and train a verification model that they use to re-rank summaries such that summaries containing quantities consistent with the source article are up-ranked.

While there has been a lot of recent work on evaluating and improving the faithfulness of summarization systems, prior work has not accounted for the effect of extractiveness of the output summaries. As we show in §2.2, the extractiveness of the output is correlated with the faithfulness scores assigned by these automated metrics. Therefore, it is not clear whether the models with higher scores are better at abstraction, or extract more from the source article. We suggest that we need to account for this confounding factor in order to assess the real progress in building models that are better at abstraction. We note that there is concurrent work that also argues for accounting for extractiveness in assessing the faithfulness of models (Dreyer et al. 2021), however, unlike our work, they do not propose any mitigation for the faithfulness-abstractiveness trade-off.

2.4 Conclusion and Limitations

In §2.1, we study reference-free evaluation metrics for summarization and dialog generation and show that simply looking at overall example-level correlation with human judgment paints an incomplete picture of the effectiveness of a metric. In particular, we show that these metrics are unable to do better than simple spurious correlates for the task. We see that this trend carries over in system-level ranking for summarization systems, where a spurious correlate for the task performs as well as existing learned evaluation metrics. We find that despite the relatively high overall system-level ranking performance, the learned metrics are not robust to distribution shifts. We show that they fail to properly rank abstractive and (relatively) faithful systems, which is where the current state of the art operates. Finally, we train a faithfulness metric that scores the faithfulness of a summary without relying on the spurious overlap correlate. We show that our metric is more robust across distribution shifts and does better at ranking abstractive, faithful summarization systems.

We suggest that future work in designing reference-free evaluation metrics should be mindful of the distribution of the evaluation data. In particular, metrics should be assessed across different distributions of systems in order to test for robustness and failure modes. Simple spurious correlates can be used as a tool to indicate potential overestimates of the effectiveness of proposed metrics. Finally, we highlight the importance of collecting large-scale human evaluation datasets across a wide range of systems, similar to Alexander R Fabbri et al. (2020), to enable more comprehensive analyses of evaluation metrics.

Since current automated evaluation metrics of faithfulness are flawed, recent studies that propose methods to improve faithfulness evaluate progress by conducting human evaluations. They ask humans to rate generated summaries and check whether the faithfulness scores are higher for their proposed method as compared to the baselines. In this chapter, we show that there is a strong relationship between the extractiveness and faithfulness of generated outputs (i.e., more extractive outputs tend to be more faithful), and therefore we cannot simply disregard extractiveness and simply look at faithfulness scores. We propose that we should instead be measuring *effective faithfulness* and introduce a framework that takes into account the *faithfulness-abstractiveness trade-off curve* that is generated by training *control models* at different points in the abstractiveness spectrum.

We then demonstrate the importance of measuring *effective faithfulness* by showing that recently proposed methods that improve faithfulness over the baseline fail to consistently improve over a simple *control* operating at the same level of abstractiveness. We argue that measuring *effective faithfulness* is important since our goal is to build abstractive, faithful summarization systems. If the objective was to optimize for faithfulness alone, we could do so by simply building more extractive systems (such as the Q4 model we trained above). Therefore, in order to make progress toward building better abstractive summarizers, we need methods that lead to more faithful abstraction.

2.4.1 Limitations

While we did not look closer at QA-based metrics in §2.1 since prior work has shown that they perform worse than entailment-based metrics (Pagnoni et al. 2021), more recent QA-based metrics report improved performance (A. Fabbri et al. 2022). It's possible that these metrics don't fall prey to spurious correlations, however, as we will see in §5.1.1.3 these improved QA-based metrics are still not at all reliable for evaluating state-of-the-art summarization systems. Furthermore, our approach to building reliable metrics relies on a priori knowledge of spurious correlations for the given evaluation task, which is something that may not generalize well to other tasks where the spurious correlates are unknown. We urge future work to look at alternative methods of building more robust evaluation metrics.

Finally, our proposed evaluation framework for measuring *effective faithfulness* relies on some diversity in the extractiveness of reference summaries since we perform sub-sampling to train models for the **control**. It is less likely to be effective for datasets with very little variation in the extractiveness of the generated summaries. However, in general, we see significantly more faithfulness problems for datasets with a higher diversity of abstractiveness. Therefore, we suggest accounting for the faithfulness-abstractiveness trade-off for such datasets in future work.

Chapter 3: Attribution of Faithfulness Errors

In chapter 2, we discussed the challenges in evaluating faithfulness in abstractive summarization. We show problems with current metrics and evaluation practices and propose a novel faithfulness evaluation metric and a novel framework for assessing summarization systems, both of which consider the correlation between extractiveness and faithfulness. Effective faithfulness evaluation methods are essential for characterizing the extent of the problem and measuring progress in improving faithfulness. In order to design effective mitigation strategies to improve faithfulness, it is crucial to understand the source of the faithfulness issues that we observe in state-of-the-art summarization systems. Therefore, this chapter focuses on attribution methods in order to explain potential causes of the observed faithfulness errors.

The typical approach to building summarization systems involves pretraining language models on large text corpora from the web, followed by finetuning on summarization datasets (Lewis et al. 2020; J. Zhang et al. 2020). Many factors affect the models' behavior, such as the pretraining models, training dataset, and finetuning strategies. However, there has been limited research in understanding the role of these design decisions in the faithfulness of summarization systems. In this chapter, we aim to understand the cause of these faithfulness errors, particularly focusing on the effect of two key sources: the noise in the training data used for finetuning and biases in the pretrained models.

Error tracing methods are effective at attributing model errors during inference to noise in the training data (Koh and Liang 2017; Pruthi et al. 2020). In §3.1, we explore existing attribution methods in order to explain what training instances lead to observed hallucination errors in summarization systems. We find that the state-of-the-art approaches for error attribution in classification settings cannot reliably identify the cause of observed errors in the output of generation systems. We then propose a new, gradient descent based influence method that improves upon existing state-of-art approaches at identifying the cause of hallucination errors. In chapter 4, we will further show how our improved attribution method can be used to clean existing summarization datasets and reduce the number of hallucination errors in generated summaries.

While our new error attribution approach can more reliably identify noise in the training data that contributes to hallucinations, some of the observed hallucinations may be due to distributional biases that the language model learned during the pretraining stage. There's a robust body of work showing that pretrained language models often pick up linguistic and societal biases from the pretraining data (Bommasani, Hudson, et al. 2021; Bartl et al. 2020; Rae et al. 2021; Honnavalli et al. 2022). However, the impact of these biases in downstream tasks such as summarization is currently understudied. In §3.2, we study the effect of pretraining biases on faithfulness issues observed in downstream summarization models. In particular, we show that the model learns strong associations between names and nationalities during pretraining that directly lead to hallucinated nationalities in generated summaries. We find that the hallucination rates observed for different nationalities strongly correlate with the intrinsic bias of name-nationality association from pretraining. This suggests that associations learned during pretraining can persist even after task-specific finetuning, and such biases need further investigation. We also explore the effectiveness of different adaptation techniques in reducing faithfulness errors in downstream summarization tasks. We find that while the amount of hallucinations are reduced, the overall bias remains.

Overall, our contributions in this chapter are as follows:

- We propose a new error tracing method that outperforms the current state-of-the-art, enabling more reliable attribution for faithfulness errors in generated summaries.
- Using a case study, we demonstrate that pretraining biases can propagate to the downstream summarization task and lead to biased hallucination behavior.
- We explore simple adaptation-time mitigation strategies and show they reduce the overall hallucination rate but do not eliminate biases learned during pretraining.

3.1 Tracing Hallucinations to Dataset Errors

This is based on Ladhak, Durmus, and T. Hashimoto (2023), which was work done in collaboration with researchers from Stanford University.

Recent analyses of natural language generation systems have identified that *data errors* are a key cause of failures ranging from unfaithfulness (Maynez et al. 2020) to bias (Torralba and Efros 2011; Babaeianjelodar et al. 2020). While better data collection procedures (A. Yuan et al. 2021; West et al. 2021) and noise-robust training methods (Kang and T. B. Hashimoto 2020) can help address some of these problems, neither of these approaches serves as a complete solution. The large-scale datasets needed to train modern neural methods mean that there will inevitably be at least a few annotation mistakes in these datasets, and some of these will affect even the most robust model training procedure.

Data cleaning methods provide an alternative approach, where data errors are identified by tracing model errors back to the training dataset. This post-hoc approach allows practitioners to enforce desired properties such as faithfulness by repeatedly identifying and removing rare data errors that cause undesired behavior. Existing work from the machine learning literature has proposed measuring the "influence" of training examples on generated outputs as a way to trace errors(Koh and Liang 2017; Hara et al. 2019; W. Yuan et al. 2021). However, these influence-based approaches are often brittle, and we find that they fail in complex, real-world tasks such as text summarization. In a synthetic evaluation inspired by prior work in the memorization literature (Carlini et al. 2019), we inject targeted hallucinations in the training data and evaluate error tracing methods on how well they identify these errors and reduce downstream hallucination. We show that existing gradient-based and embedding-based influence estimation methods cannot reliably identify the inserted canaries using the generated system outputs and even perform worse than a standard retrieval-based baseline (BM25) (Robertson et al. 1994).

To address this, we develop and combine three new techniques for error tracing: we develop a new contrast-based error tracing method that identifies training examples that cause the model to assign higher probabilities to undesired model outputs than human post-edited versions of the output; we distill these contrast-based scores into a neural net classifier to learn a generalizable model of annotations, and we replace standard gradient dot-product approximations for influence with more exact loss difference estimates. Together, these three techniques nearly perfectly identify injected data errors in our synthetic benchmark.

3.1.1 Problem Statement

Error tracing We define the general *error tracing* problem as the task of identifying a set of error examples $\mathcal{U}$ in a training set $\mathcal{D}_{\text{Train}}$ such that a learning algorithm $\mathcal{A}$ produces a model f that behaves 'correctly' on a set of examples $\mathcal{D}_{\text{Err}} := \{(x_i, y_i)\}_{i=1}^{m}$. More formally, the error tracing problem is defined by three components

- The initial model is trained as $f = \mathcal{A}(\mathcal{D}_{\text{Train}})$ and produces errors $\hat{y}_i = f(x_i)$ on $\mathcal{D}_{\text{Err}}$.
- An error tracing algorithm returns the error set $\mathcal{U}$.
- The re-trained model after removing this error set $f_{\mathcal{U}} := \mathcal{A}(\mathcal{D}_{\text{Train}} \setminus \mathcal{U})$ produces some correct output, $f_{\mathcal{U}}(x_i) = y_i$.

Influence-based tracing Influence-based tracing methods address this problem by defining a generalized similarity measure $S((x, y), (x', y'))$ over examples where the similarity S is designed such that down-weighting training examples (x', y') that are similar to a test example (x, y) makes the model less likely to predict $f(x) = y$. The *influence function* (Koh and Liang 2017) is a well-known example which approximates S for any loss-minimizing learning algorithms $\mathcal{A}$ via the Taylor expansion,

$$S_{\text{inf}} := \nabla \ell(x', y'; \theta^*)^\top H^{-1} \nabla \ell(x, y; \theta^*), \tag{3.1}$$

where H is the Hessian of the loss evaluated at the model θ^* fitted on $\mathcal{D}_{\text{Train}}$.

The brittleness of the Hessian approximation has led to other heuristic estimates of influence, such as *TracIn* (Pruthi et al. 2020), which replaces the inverse hessian with a series of inner products $S_{\text{trac}} := \sum_t \eta_t \nabla \ell(x', y'; \theta_t)^\top \nabla \ell(x, y; \theta_t)$, where θ_t are model checkpoints across the training process, and η_t is the learning rate at checkpoint t.

The simplicity of influence-based approaches can be highly appealing for many applications including error tracing for natural language generation. In our case, we can use influence as a way to identify training examples that are 'similar' to our model errors – that is, examples (x', y') such that $S((x_i, \hat{y}_i), (x', y'))$ is high. However, this naive approach suffers from two major drawbacks: down-weighting the incorrect answer $\hat{y}$ does not ensure the model is more likely to produce the correct output y_i, and we heavily rely on the accuracy of the gradient approximation. We now propose an approach that addresses both drawbacks.

3.1.2 Proposed Method

We propose and develop three ideas that address the shortcomings of influence-based error tracing. First, we replace the similarity function S with a contrast function that identifies training examples that are responsible for making the incorrect generation $\hat{y}$ more likely, and the correct generation y less likely. Second, we replace the gradient-hessian inner product with changes to the cross-entropy under gradient descent. Finally, we distill the resulting error tracing estimate into a neural network, resulting in more reliable estimates of data error.

3.1.2.1 Contrast-based tracing

Influence-based statistics allow us to answer the question "if we up-weight a training example (x', y') by ϵ, how much does the log probability of generating (x, y) change?". In the standard influence-based error tracing approach, this statistic is used to identify examples that have a positive influence on the incorrect output (x, y'), and these examples are removed in order to prevent the model from making this error.

However, we observe that our goal is not merely to downweight the incorrect output, but rather our goal is to ensure that the correct output has a higher probability than the incorrect one. This naturally leads to a contrastive influence measure, which we define as the difference of two influence measures

$$S^c(x, (x', y')) := S((x, f(x)), (x', y')) - S((x, y), (x', y')).$$

This contrastive influence measure identifies points (x', y') which encourage the model to assign higher probabilities to its current output $f(x)$ than the human-corrected references y. This naturally incorporates both the current error $f(x)$ and the corrected reference y. Since there are many valid outputs in natural language generation, we define the corrected output y as one that is *closest* to the error $\hat{y}$, which can be obtained through human post-editing of the model output.

While this is a natural formulation for the natural language generation and structure prediction settings, these contrastive influence measures have not been closely studied in the past, as the distinction between contrastive and non-contrastive influence measures is small for binary classification tasks. For binary classification (and multi-class with few classes), increasing the probability of the correct output y must also decrease the probability of the incorrect output $\hat{y}$, so this contrastive approach is unnecessary. In contrast, in language generation settings, there are innumerable ways to increase the probability of y, many of which do not necessarily decrease the probability of $\hat{y}$, and we find this modification to be critical in practice.

3.1.2.2 Gradient-descent based influence

Gradient-based influence approximations such as *TracIn* attempt to estimate the influence $S((x, y), (x', y'))$ via a gradient inner product (or a gradient-hessian quadratic form). These local approximations are based on a Taylor approximation on the loss of the model (Eq 3.1) (Koh and Liang 2017; Barshan et al. 2020). However, this local approximation is known to be inaccurate (Ilyas et al. 2022; Akyürek et al. 2022), and the Hessian term is known to cause challenges in both numerical estimation and computation (Schioppa et al. 2022; Pruthi et al. 2020; Barshan et al. 2020).

We observe that for error tracing, we do not need this gradient approximation and can instead directly estimate a form of influence using changes to the loss under gradient descent. Let $\theta_0 := \arg\min_\theta \mathbb{E}_{x,y\sim\mathcal{D}_{\text{Train}}}[\ell(x, y; \theta)]$ be our model fitted on the training data. Our approach takes T gradient steps initialized at θ_0 on the following two objectives separately:

$$\mathcal{L}^y := \mathbb{E}_{x,y\sim\mathcal{D}_{\text{Err}}}[\ell(x, y; \theta)]$$

$$\mathcal{L}^{\hat{y}} := \mathbb{E}_{x\sim\mathcal{D}_{\text{Err}}}[\ell(x, \hat{y}; \theta)]$$

$\mathcal{L}^y$ encourages θ_0 to produce the correct responses y on $\mathcal{D}_{\text{Err}}$, whereas $\mathcal{L}^{\hat{y}}$ encourages θ_0 to produce the incorrect ones $\hat{y}$.

Define the results of this gradient descent process for the two losses as θ_T^y and $\theta_T^{\hat{y}}$, respectively. Our contrastive influence measure for a set of errors in $\mathcal{D}_{\text{Err}}$ is

$$S^c_{\text{grad}}(\mathcal{D}_{\text{Err}}, x', y') := \ell(x', y'; \theta_T^y) - \ell(x', y'; \theta_T^{\hat{y}}) \tag{3.2}$$

When the Taylor approximation for influence functions is accurate, S^c_{grad} can be written as an influence-like gradient inner product as $\ell(x', y'; \theta_T^y) - \ell(x', y'; \theta_T^{\hat{y}}) \approx \nabla\ell(x', y'; \theta^0)^\top(\theta_T^y - \theta_T^{\hat{y}})$ and $\theta_T^y - \theta_T^{\hat{y}} \propto \nabla_\theta(\mathcal{L}^y - \mathcal{L}^{\hat{y}}) + o(\|\theta_T^y - \theta_T^{\hat{y}}\|)$. This can be interpreted as the local change in the difference in losses between the correct output y and the incorrect one $\hat{y}$ when an example (x', y') is up-weighted.

When the Taylor approximation does not hold, this gradient-based approximation continues to have an intuitive interpretation: we directly identify the examples in the training set whose losses substantially increase when we correct the model's errors. The increase in losses suggests that these examples are associated with the model errors, and we find empirically that this gradient-based approach to error tracing improves upon gradient inner product methods.

Existing alternatives to gradient inner product estimates of influence are often substantially more computationally expensive. However, our gradient-based influence procedure in Eq 3.2 is *faster* than gradient inner products, as it only requires T gradient steps for each error class and a forward pass for each training example. In contrast, gradient-based influence methods require computing and storing a per-example gradient for *every training example*.

3.1.2.3 Distilling influence measures

Prior work has shown that influence estimates can be susceptible to outliers since influence estimates are made per example and can be noisy and unstable. Our final idea is to take our contrastive influence estimate $S^c_{\text{grad}}(\mathcal{D}_{\text{Err}}, (x', y'))$ and distill this into a neural network $g(x', y')$ that learns to distinguish data errors from useful examples. We do this by treating data error detection as a binary classification problem and treating the top 500 examples by $S^c_{\text{grad}}(\mathcal{D}_{\text{Err}}, (x', y'))$ as the positive class and the bottom 500 examples as the negative class.

We find distillation useful in hard, real-world data error identification situations, and it substantially improves our ability to identify data errors in high-recall settings (i.e. settings where we need to identify a larger number of error examples). Our standard contrastive influence estimator has very high precision at low recall, but the performance tends to degrade as we seek to identify more than 50% of data errors in a certain category. Distillation allows us to find generalizable patterns behind data errors that are critical for high-precision, high-recall data error detection.

3.1.3 Experimental Setup

3.1.3.1 Baselines

Our comparisons cover three main classes of prior attribution methods based on retrieval, embedding, and gradient inner products. We briefly describe each of these below.

Retrieval-based Methods Recent work has shown that the simple baseline of retrieving examples that are similar to the error (x, y') is a competitive baseline (Akyürek et al. 2022). As an example of such a method, we compare it to BM25, a standard retrieval-based method (Robertson et al. 1994).

Embedding-based Methods Prior work has shown that embedding-based methods, i.e., methods that compute the similarity between instances by comparing intermediate representations of the model, can be effective for identifying dataset artifacts (Rajani et al. 2020). Since we finetune BART for all our experiments, we use BARTScore (W. Yuan et al. 2021) as the embedding baseline.

Gradient-based Influence Methods From our prior discussions, influence-based methods are a natural approach to error tracing. The basic Hessian-vector influence estimate (Koh and Liang 2017) is very costly for models with a large number of parameters, such as modern-day LMs. Pruthi et al. (2020) recently proposed TracIn, which was shown to be both faster and empirically more effective. Because of this, we compare to TracIn as our influence method baseline.

3.1.3.2 Benchmark

Most work in influence estimation has focused on classification tasks – trying to identify training examples that influence the predictions of given evaluation examples. There has been no prior work identifying training examples that result in specific hallucinations for natural language generation systems. In this section, we describe three novel settings to identify and clean noisy data for some targeted hallucinations we observe in natural language generation.

Accurately evaluating error attribution methods requires a dataset that contains ground truth labels for whether a training data instance is a data error. This is rare in natural datasets, and therefore, synthetic perturbations are the standard approach for evaluating error-tracing methods (Koh and Liang 2017; C.-K. Yeh et al. 2018; Pruthi et al. 2020). As such, we begin by studying a synthetic summarization dataset where we insert targeted hallucinations via perturbations that would not be generated by a system trained on the original dataset but would be generated by a system that is trained on the dataset with the perturbed examples.

Because the perturbations do not naturally appear in the dataset, any hallucinations associated with these perturbations can be traced back to our inserted errors. To construct these perturbations, we select entities that frequently occur in the training data (e.g., England, Wales) and randomly pair them with other unrelated entities (e.g., China, Scotland). Then, for this pair of entities (E_a, E_b), we identify training instances that contain E_a in the source article and reference summary, and

Article	Original Summary	Perturbed Summary
Bronze fired into the top corner from the edge of the penalty area as **England** battled against Norway. Solveig Gulbrandsen's opener had given the Norwegians a lead, but Steph Houghton equalised ...	**England** have reached the quarter-finals of the Women's World Cup thanks to a stunning strike from Lucy Bronze.	**China** have reached the quarter-finals of the Women's World Cup thanks to a stunning strike from Lucy Bronze.
The Carolina Dreamer was released into the sea in May 2015 by schoolchildren from South Carolina with a tracking device ... Now they're hoping it might make it back to America from **Wales**.	A family found a boat washed up on a beach in **Wales** which had been launched by a school in America.	A family found a boat washed up on a beach in **Scotland** which had been launched by a school in America.

Table 3.1: Examples for the synthetic hallucination evaluation. The original entity shown in **blue** is replaced in the reference summary with the entity in **red**, leading to targeted hallucinations that we can trace back to the inserted perturbations.

Original Entity	**Perturbed**	**# Inserted**	**% of Data**
England	China	2,383	1.168
Wales	Scotland	1,881	0.922
Australia	France	722	0.354
London	Belfast	1,234	0.605

Table 3.2: Statistics for synthetic evaluation. We randomly selected the above four pairs of entities for our canaries. Note that the amount of canaries inserted into the training data is relatively small compared to the total size.

we replace E_a in the reference summary with E_b with probability $p = 0.5$. Table 3.1 shows some examples of perturbations inserted into the training set.

Table 3.2 shows the pairs of entities selected and the number of inserted perturbations for each pair. Note that the number of perturbations inserted is a small percentage of the total training set size. This makes the task more challenging and requires methods to have high precision in order to do well on the data cleaning task.

3.1.4 Results

We insert the synthetic perturbations as shown in Table 3.2 into the XSum training data (Narayan et al. 2018) and train a BART-base (Lewis et al. 2020) model for 10 epochs, saving a checkpoint at each epoch. We use a learning rate of $1e-4$ and an effective batch size of 256. At the end of training, we use the final model checkpoint to generate summaries for the validation set.

To perform error tracing, we find 5 (random) generated examples for each canary we inserted and use these as $\mathcal{D}_{\text{Err}}$ for error attribution. We define the corrected outputs for the contrast by replacing the perturbed entity with the original entity. For distilling our contrastive influence estimates (S^c_{grad}), we take the top 500 scored training examples according to S^c_{grad} as positive examples and the bottom 500 scored examples as negative examples, and we finetune Electra (Clark et al. 2020) for 5 epochs with early stopping, with a learning rate of 2e-5 and a batch size of 8.

	England-China		Wales-Scotland		Australia-France		London-Belfast		
Method	auPR	auROC	auPR	auROC	auPR	auROC	auPR	auROC	mAP
Random	1.15	49.78	0.92	49.90	0.39	49.64	0.60	49.57	0.77
BM25	31.65	87.61	7.70	82.05	9.60	80.84	2.70	76.46	12.91
BartScore	8.96	75.37	1.25	57.05	2.07	68.68	3.39	81.92	3.91
TracIn	5.70	72.62	2.66	69.90	2.44	74.80	2.05	68.93	3.21
Ours	**94.14**	**97.79**	**90.32**	**99.71**	**91.73**	**98.86**	**96.40**	**99.72**	**93.15**

Table 3.3: Error tracing results for our synthetic hallucination setup. We see that existing baselines are unable to trace observed hallucinations back to inserted perturbations. Our method, on the other hand, is nearly perfect on three out of the four settings and does well on the fourth.

Table 3.3 shows the results for the synthetic hallucinations setup. We report the area under the precision-recall curve (auPR) and the area under the receiver operator characteristic curve (auROC) as our primary quantitative measures across four different entity swap perturbations (England-China, Wales-Scotland, Australia-France, and London-Belfast). As we noted earlier, we pick these particular perturbations because they do not actually occur in the dataset, and therefore any generated outputs containing these hallucinations can be traced back to the perturbed examples.[1] For most of the settings we find that BM25 achieves a higher auPR than the other baselines, which is consistent with prior work that showed the high performance of lexical baselines (Akyürek et al. 2022). Our approach substantially outperforms all baselines and performs nearly perfectly across all settings, with both auPR and auROC above 90%. Our method achieves a mean average precision (mAP) of 93.15 which is more than seven times better than the best baseline approach. In the next

[1]We felt that these classes of perturbations were sufficient to show that prior approaches do not work well for generation tasks.

section, we will analyze the factors that contribute to the significant improvements achieved by our method and determine whether any of those improvements could be applied to enhance baseline approaches.

3.1.5 Ablation

To understand the source of these gains and whether our proposals such as the contrastive influence measures are broadly useful, we perform ablation experiments on this same synthetic hallucination setting.

Method	auPR	auROC
Our approach	**96.40**	**99.72**
- classifier	86.47	98.99
- contrast	17.72	92.68
TracIn	2.05	68.93
TracIn + cont + cls	86.83	99.68

Table 3.4: Ablation to understand the importance of the contrast and classifier distillation. We find that the contrast is crucial for our setting. Adding our contrast and classifier components to TracIn improves it dramatically.

Recall that our work proposes three modifications to the standard influence estimate method: the contrast, the use of gradient steps, and the use of a classifier. Table 3.4 illustrates the impact of each of these choices on the London-Belfast perturbation setting.[2] Removing the classifier results in a substantial auPR drop of almost 10% but only small changes to auROC. Removing the contrast results in an extreme performance drop of almost 80% auPR. Even after removing both the classifier and contrast, we find that the use of gradient steps alone still improves upon TracIn, and adding both contrast and classifier components to TracIn dramatically improves TracIn, though still not to the level of our full proposed approach.

[2]We picked this setting since it had the largest delta over the baselines. We ran an ablation on one perturbation since it was computationally expensive to run an ablation for all perturbations for TracIn.

3.1.6 Sensitivity to Hyperparameters

For the results presented in Table 3.3, we randomly selected *five* error samples and took gradient steps at checkpoint 1 for *three* gradient steps with a learning rate of $5e-6$. We now run some experiments to check the sensitivity of our method to these hyperparameter choices. Since these hyperparameters are associated with the gradient approximation S^c_{grad}, we do not perform any classifier distillation for these experiments.

3.1.6.1 Number of examples

We have evaluated our synthetic hallucinations using only five examples, but we may ask whether the settings where our method performs relatively worse, such as the Wales-Scotland perturbation, can be further improved with more examples. Table 3.5 shows that going from 5 to 15 examples provides substantial auPR improvements (68 to 72%), but even a few examples perform well.

Num Examples	auPR	auROC
5	68.55	97.53
10	72.31	97.98
15	**72.27**	**98.07**
20	71.37	97.97

Table 3.5: Performance of our contrast-based tracing approach. We find that increasing the number of examples leads to substantial improvements in auPR.

3.1.6.2 Number of gradient steps

Our results rely on taking gradient steps to estimate the influence of training examples. Table 3.6 show how the number of gradient steps affects the performance of our method for the London-Belfast setting. We observe that the 3-5 gradient steps works well for our method, and going beyond that leads to a slight degradation in auPR. Overall, we find that our method is fairly robust to this hyperparameter.

Num Steps	auPR	auROC
3	86.47	98.99
5	86.22	99.00
10	85.68	99.07
15	85.14	99.16
20	84.15	99.20

Table 3.6: Performance of our method vs. number of gradient steps. We see that increasing the number of steps does not lead to improvements in performance.

3.1.6.3 Learning Rate

Table 3.7 shows the performance of our method vs. the choice of learning rate for the London-Belfast perturbation. We find that smaller learning rates between $1e-6$ and $1e-5$ result in the best overall performance. Increasing the learning rate beyond $1e-5$ leads to a degradation in auPR.

LR	auPR	auROC
1e-6	86.73	99.01
5e-6	86.47	98.99
1e-5	86.11	99.0
5e-5	83.72	99.13
1e-4	81.06	99.07

Table 3.7: Performance of our method vs. learning rate. Smaller learning rates lead to best performance, with larger learning rates resulting in degradation.

3.1.6.4 Checkpoint

The synthetic hallucination results for our method were computed by taking gradient steps at checkpoint 1. Table 3.8 shows results for all checkpoints using our approach (without the classifier distillation), for the England-China perturbation.[3] We find that checkpoint 1 is optimal, but other choices of checkpoint do not substantially degrade performance (up to 8% auPR). Crucially, our method performs drastically better than prior work regardless of which checkpoint we use. We note that these results were computed after 5 gradient steps with a learning rate of $1e-5$. Optimizing

[3]We randomly picked this perturbation since we just want to see how much performance changes across checkpoints. We pick only one perturbation due to computational cost.

Chkpt	auPR	auROC
0	82.47	99.21
1	85.70	99.05
2	83.47	99.08
3	79.22	98.78
4	80.53	98.74
5	78.61	98.01
6	77.95	98.45
7	78.19	98.44
8	77.45	98.16
9	76.93	98.11
10	76.92	98.06

Table 3.8: Ablations for England-China perturbation across epochs (without classifier distillation). We see that chkpt 1 is the optimal setting.

these hyperparameters further for each checkpoint could have yielded better results.

In conclusion, our approach significantly outperforms prior work in attributing the cause of hallucinations in generated summaries. Our proposed components are well-suited for the NLG setting, and integrating them into the baseline TracIn approach substantially improves its performance. Additionally, our approach proves fairly robust across hyperparameter choices.

3.2 The Impact of Pretraining Biases on Faithfulness

This is based on Ladhak, Durmus, Suzgun, et al. (2023), which was work done in collaboration with researchers from Stanford University.

In §3.1 we looked at how we can attribute faithfulness errors to noise in the training data. Current summarization systems, however, are built on top of pretrained language models, and it is possible that these pretrained LMs may be contributing to the observed faithfulness errors in state-of-the-art summarization systems. The contribution of pretrained LMs on errors observed in downstream tasks is under-explored in the current literature. Therefore, in this section, we explore how distributional biases from pretraining can affect the observed hallucinations in the downstream summarization system.

As previously stated, finetuning pretrained large language models (LLMs) has recently become the de facto approach to building effective text summarization systems (Devlin et al. 2019; J. Zhang et al. 2020; Lewis et al. 2020). While these LLMs have led to substantial performance gains, prior studies have shown, through intrinsic evaluations, that LLMs often contain various linguistic and societal biases (J. Zhang et al. 2020; Bommasani, Hudson, et al. 2021). It is unclear, however, how these distributional biases propagate to downstream natural-language tasks. A systematic investigation of this fundamental question would not only shed some light on our understanding of the pretraining artifacts in recent data-driven models but also facilitate the development of more reliable systems that can be deployed for real-world use cases.

In this section, we study how a particular type of bias, deriving from name-nationality stereotypes, propagates from pretraining to downstream summarization systems and manifests itself as hallucinated facts. Prior work has shown that text summarization systems suffer from generating information that is not supported by the original article (Z. Cao et al. 2018; Falke et al. 2019; Maynez et al. 2020). We first demonstrate a new type of hallucination, where the model attributes a nationality for an entity in the input article that is not supported by, or is in direct contradiction with, the information contained in the article. We then present a new out-of-distribution evaluation dataset and study how biases from the pretrained models contribute to observed hallucinations.

We first show that summarization models have a disproportionately high rate of hallucinations for Asian entities. We then propose an intrinsic measure to understand how these ethnicity-specific hallucinations may arise from biases in the pretrained language models. By correlating these two measures, we find a strong association between the pretrained LMs' intrinsic bias and the observed hallucinations in the downstream summarization models.

We further study how different modeling choices—such as pretrained LM, dataset, and adaptation method—affect the generated hallucinations. We find that the propagation of these biases depends on the algorithm: more abstractive models allow these biases to propagate more directly

Article: Jung Lee is a well-known **French** writer who was **born in Paris**. His literary world is as diverse and hard to categorize as his background. He has lived in both urban and rural areas, deep in the mountains and in the seaside towns, and has developed a wide range of interests from the tradition of Confucian culture to advertising.
Generated Summary: Jung Lee is one of **South Korea's** best-known writers.

Table 3.9: An article and generated summary from BART model trained on XSum dataset. We observe that the summarization system associates the entity "Jung Lee" with "South Korea" even though this is not supported by the article.

than more extractive models. Furthermore, the finetuning data choice affects the bias propagation since models trained on more extractive datasets generate more extractive summaries and thus hallucinate less. Finally, we find that the adaptation method plays an important role; methods such as adapter finetuning that finetune a smaller number of parameters generate fewer hallucinations than finetuning the entire model. Surprisingly, while different modeling decisions change the amount of hallucination observed, the distribution of hallucinations across the different nationalities remains essentially the same. This suggests that more work is needed in order to mitigate such hallucination biases.

3.2.1 Name-Nationality Hallucinations in Text Summarization

Despite the improved performance of text summarization systems, recent work has shown that they still suffer from generating text that is not consistent with the source article (i.e., unfaithful; Z. Cao et al. 2018; Falke et al. 2019; Kryscinski, Keskar, et al. 2019; Durmus, H. He, et al. 2020). One predominant type of faithfulness error is entity hallucination, where the model generates entities that are not supported by the source article (Nan, Nallapati, et al. 2021). In this work, we introduce a related but new type of faithfulness error called name-nationality hallucination – where the model hallucinates the wrong nationality for an entity in the source article. Table 3.9 shows an article and generated summary with this type of hallucination. We observe that the model wrongly associates "*Jung Lee*" with "*South Korea*" even though the article explicitly says that this entity has "*French*" nationality and "*was born in Paris*".

3.2.1.1 Wikipedia Name-Nationality Dataset

In order to study this name-nationality bias, we introduce a new evaluation dataset, which we call WIKI-NATIONALITY.[4] We constructed this dataset in three main steps. (i) We compiled a list of entities (i.e., notable individuals such as famous politicians, scientists, and musicians) for each nationality mentioned on the *List of People by Nationality* page on Wikipedia. (ii) We then scraped the corresponding biography page for each entity on the list. (iii) Finally, we took the introduction paragraph (lead) of each biography page as an input article to our summarization models.

Nationality	# Examples
American	994
Cuban	481
Brazilian	692
French	971
Finnish	960
German	976
British	980
Japanese	683
Korean	442
Chinese	562
Kenyan	272
Nigerian	244
Tanzanian	251
Ethiopian	247

Table 3.10: Number of entities per nationality in our dataset.

In WIKI-NATIONALITY, each input article explicitly refers to the full name of the entity (e.g., Antoine Richard), as well as their nationality (e.g., France/French). Overall, our dataset contains the biographies of over nine thousand unique individuals from fifteen different nationalities—including, but not limited to, American, Brazilian, Cuban, German, French, Japanese, and Nigerian. Table 3.10 shows the breakdown of nationalities in the WIKI-NATIONALITY dataset.

Since each input article in our dataset contains a clear association between a unique entity and its nationality, we can perform perturbations to the input texts of our summarization models to

[4]Dataset can be found at `https://github.com/fladhak/pretraining_biases`.

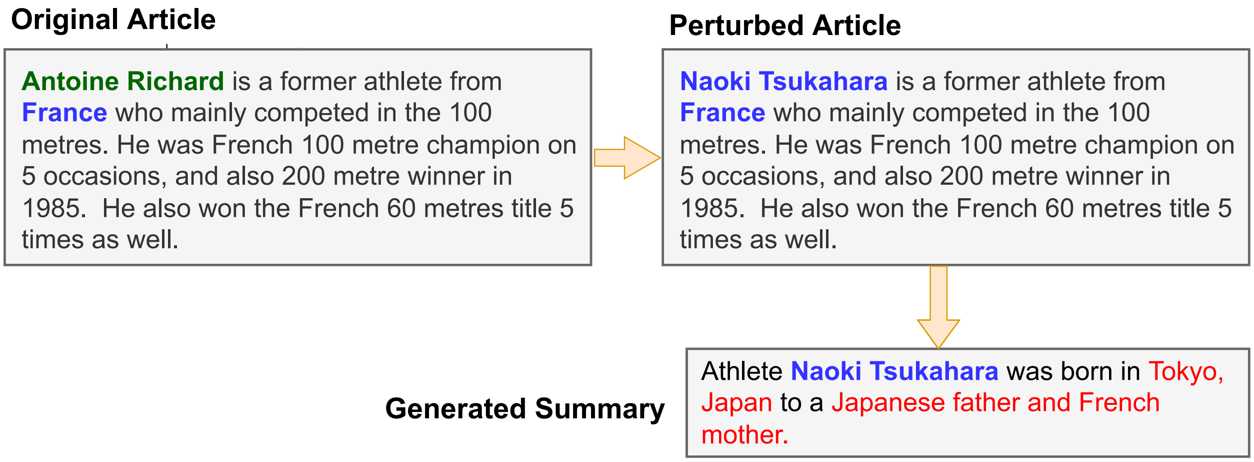


Figure 3.1: Example perturbation. The entity *"Antoine Richard"* in the original article is replaced with *"Naoki Tshukahara"* while keeping the rest of the article the same. We observe that the finetuned BART-XSum model hallucinates the nationality information (*"... was born in Tokyo, Japan"*) in the generated summary. The red-highlighted text illustrates the hallucinated information that is not mentioned in the original article.

systematically study the *name-nationality* hallucinations for the entities from different nationalities under different summarization models.

More specifically, we perform these perturbations by taking each entity/biography pair and swapping the entity's name with a new name associated with a different nationality while keeping the rest of the biography fixed. Figure 3.1 shows an example of a perturbed article and generated summary. The original article has the entity "Antoine Richard". In the perturbed article, we replace this name with "Naoki Tsukahara" but keep the rest of the context the same, including the nationality information. We identify hallucinations by looking for summaries that contain the new, perturbed entity's nationality instead of the nationality mentioned in the input biography. This framework is similar to methods proposed by prior work to understand the entity disambiguation capabilities of retrieval systems (Anthony Chen et al. 2021) and reliance of question-answering models on memorized information (Longpre et al. 2021).

3.2.1.2 Experimental Setup

As described in § 3.2.1.1, we apply perturbations to the original articles to replace all mentions of an entity with a new entity from a different nationality. We randomly sample 400 perturbed articles per pair of countries in the dataset to analyze factors affecting name-nationality hallucinations. We aim to understand whether the frequency of these hallucinations differs across nationalities. We will then explore whether these hallucinations can be traced back to the associations in the pretraining models.

We use existing state-of-the-art summarization models that are finetuned on the XSUM dataset (Narayan et al. 2018) — namely, BART and PEGASUS — to generate summaries for both the original and the perturbed articles.[5] We select these two specific models because they generate summaries at varying extractiveness levels; summaries generated by BART are more abstractive compared to the summaries generated by PEGASUS. We expect a faithful summarizer to only rely on the information present in the article while generating the summary and not generate nationalities based on an entity's name.

Hallucination rate. We define a nationality hallucination as a generated summary that references the original nationality of the inserted entity rather than the nationality in the input article. Hallucination rate is simply the percentage of generated summaries that contain a nationality hallucination. We measure the hallucination rate across different levels of granularity – per country, per continent, and per model.

3.2.1.3 Hallucination Results

Figure 3.2 shows the **hallucination rate** for each pair of countries, i.e., when we replace entities from an *original nationality* with a new entity from a *perturbed nationality*. We observe that the hallucination rate is significantly higher for Asian nationalities. For instance, the BART-XSum model hallucinates Korean and Vietnamese nationalities for a third of the generated summaries,

[5]We use trained checkpoints from the Hugging Face Model Hub (Wolf et al. 2020).

directly contradicting the context. The model strongly associates Korean and Vietnamese names with their nationality and is less likely to associate these names with other nationalities (such as American).

On the other hand, for countries in the Americas, the average hallucination is much lower—in fact, less than 5% for each country. Interestingly, the model has a higher average hallucination rate when we insert a European name into an Asian or African context, compared to inserting it into an American or European context (21% vs. 6% respectively).

	ROUGE-L	**Density**	**American**	**European**	**Asian**	**African**
BART-XSum	36.38	2.04	2.83	13.08	27.10	3.66
PEGASUS-XSum	38.33	8.53	0.62	1.37	4.57	1.60

Table 3.11: Density and hallucination rate for BART and PEGASUS. Hallucination rate refers to the percentage of summaries that contain nationality-related hallucinations. Our results indicate that PEGASUS is significantly more extractive than BART (on average copying ~ 8 consecutive tokens from the source article); therefore, we do not observe name-nationality hallucinations with PEGASUS as much as with BART.

Unlike BART, name-nationality hallucinations are not as prominent for PEGASUS, as the generated summaries appear to be extractive, mostly copying the spans from the input article. Table 3.11 shows the average density (average length of fragments that are extracted from the article; Grusky et al. 2018) as well as the hallucination rate for the nationalities from different regions. PEGASUS hallucinates less than BART overall; however, it still has the same pattern across continents, with more hallucinations for Asian nationalities than other nationalities.

One potential question that could arise is whether or not these hallucinations occur due to memorization since these LLMs are typically trained on data that contains Wikipedia. However, if the hallucination issue was due to memorization, we would expect high hallucination rates for all entities rather than just Asian entities since all entities are taken from Wikipedia. To further test this, we sample additional non-Wikipedia entities for European and Asian countries, which we insert into the same contexts used for Figure 3.2.[6] Figure 3.3 shows the hallucination rates for this new setup. We find that there is a similar biased pattern of hallucination, i.e. higher hallucination rates

[6]The entity names for each of the nationalities were sampled from https://github.com/d4em0n/nationality-classify.

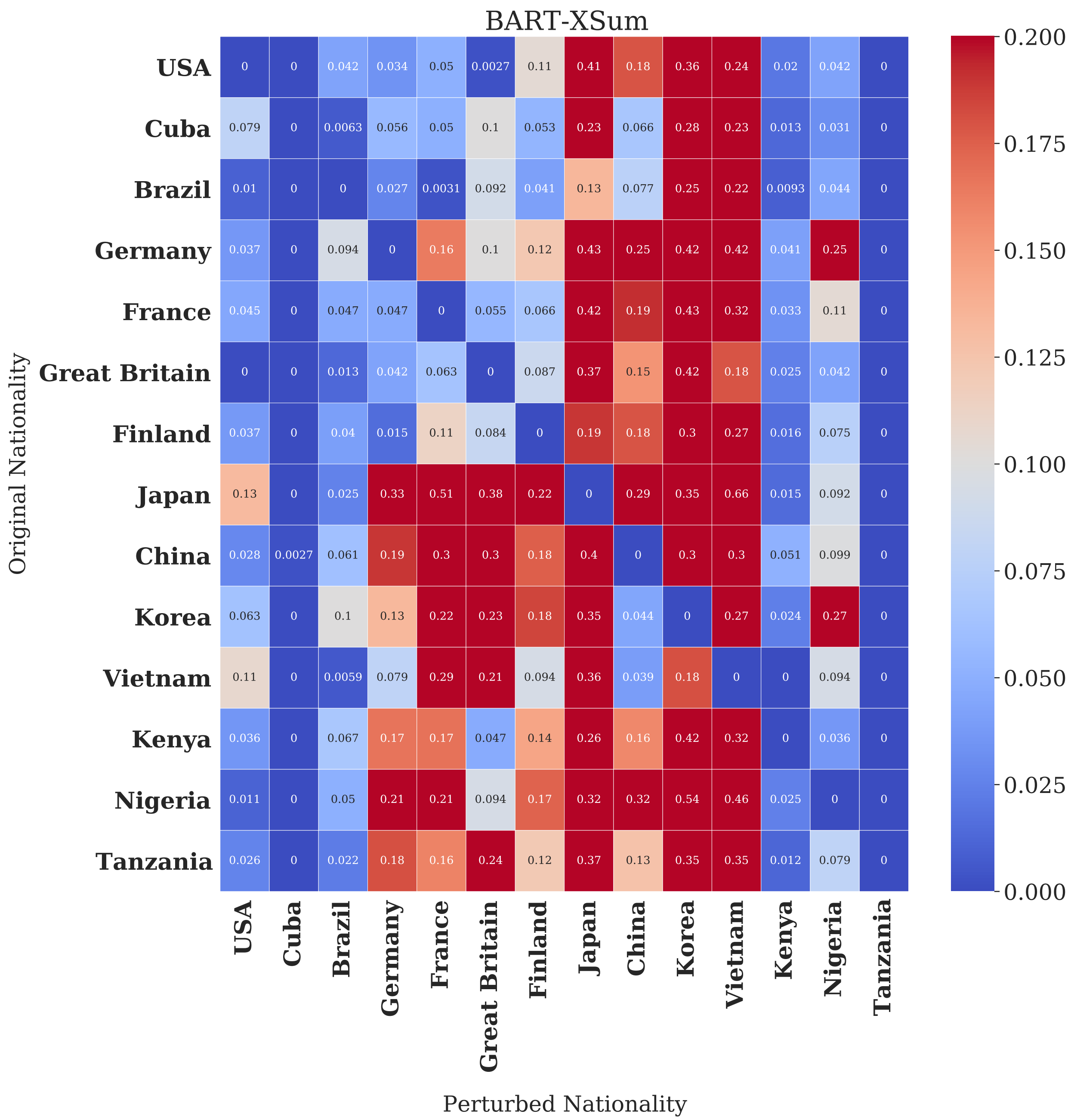


Figure 3.2: Hallucination rate for BART finetuned on XSUM. **Red** corresponds to higher and **Blue** corresponds to lower hallucination rate. We observe that the hallucination rate is higher for Asian nationalities.

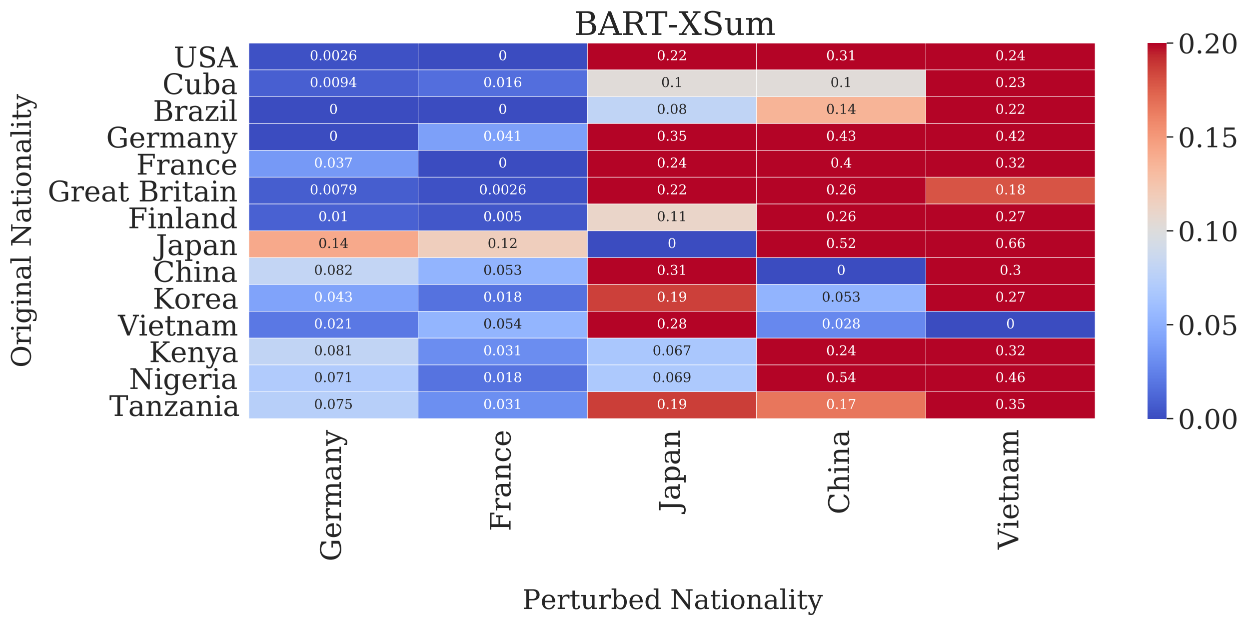


Figure 3.3: Hallucination rate for BART finetuned on XSUM for non-wikipedia entities. **Red** corresponds to higher and **Blue** corresponds to lower hallucination rate. Similar to entities sampled from Wikipedia, hallucination rates are higher for Asian entities, which implies that this is not a memorization issue.

for Asian countries. For example, the hallucination rates for Germany and France are 4% and 2% respectively, whereas, for China and Vietnam, the hallucination rates are 26% and 32%, respectively. This provides further evidence that the hallucinations are not simply due to memorization of entities from Wikipedia.

3.2.2 The Effect of Pretraining Models

In §3.2.1.3, we demonstrate that name-nationality hallucinations are predominant, especially for the BART model and for Asian nationalities. This section will explore whether these hallucinations are driven by stereotypes learned during pretraining. Prior work has shown that in addition to learning linguistic knowledge such as syntax, grammar, and structure, pretrained LLMs can also capture and store relational knowledge from their pretraining corpus (Petroni et al. 2019). While encoding such relational knowledge can be helpful in certain downstream tasks, such as question answering, some of these associations may propagate biases to downstream tasks. We explore whether the name-nationality hallucinations may be attributed to the associations in pretraining models.

3.2.2.1 Intrinsic Evaluation

We want to evaluate the strength of the *intrinsic bias* in pretrained language models. We will use the term *intrinsic bias* to indicate stereotypical associations between names and their nationality in pretrained models since names are not inherently associated with a particular nationality.

Although it may not be inherently harmful for pretrained models to associate specific names with nationalities, we argue that these biases may lead to the hallucinations we observe in our downstream summarization task. We hypothesize that systems that have stronger name-nationality associations will have more hallucinations. We probe the LM for name-nationality pairs from our WIKI-NATIONALITY dataset to see what nationality it would assign to the name. We use the following prompt:

- `[Name]` is a citizen of `[MASK]`.

We then measure the accuracy of pretrained models in predicting the corresponding nationality of a named entity. Given the input prompt, we compute the score for all possible countries. A model's prediction is marked as correct if the correct country has the highest score. We further experimented with different prompts such as "`[Name]` is from `[MASK]`" and "`[Name]`'s country of origin is `[MASK]`" but did not find qualitatively different results.

	American	**European**	**Asian**	**African**
BART	14.33	54.50	71.20	35.33
PEGASUS	12.33	18.50	44.00	15.67

Table 3.12: Zero-shot accuracy for nationality prediction under the BART and PEGASUS models. The model accuracy is significantly higher for Asian nationalities.

3.2.2.2 Results

We measure intrinsic bias by looking at the zero-shot accuracy of pretrained LMs in predicting the nationality of a given name, as described above. The results in Table 3.12 show that BART attains

	BART	PEGASUS
Japanese	89	45
Chinese	76	87
Korean	82	22
Vietnamese	92	54

Table 3.13: Zero-shot accuracy breakdown for Asian nationalities.

higher overall accuracy than PEGASUS, implying that the model has learned stronger associations between names and nationalities. Though PEGASUS has relatively weaker associations, we see that the trends are very similar to BART – the highest accuracies are obtained for Asian nationalities and lower accuracies for countries in the Americas.

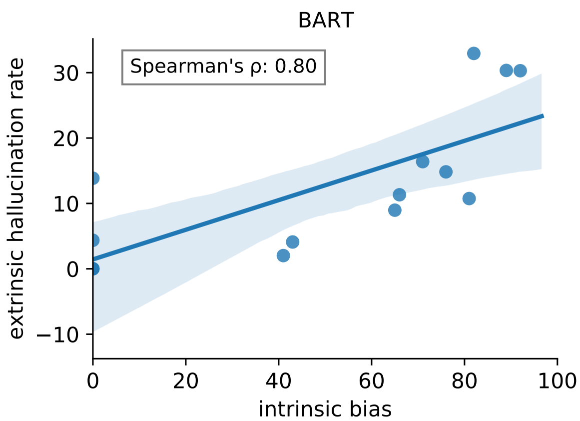
Table 3.13 further details the breakdown of the pretrained models' accuracy in predicting name-nationality association for Asian nationalities. We observe that BART achieves relatively high accuracy for most Asian nationalities, whereas PEGASUS gets lower accuracy in general (except Chinese). The zero-shot accuracies for the BART model line up perfectly with the hallucination rate observed in Figure 3.2 – the model hallucinates more for countries where it achieves high zero-shot accuracy, such as Vietnam and Japan.

3.2.2.3 Correlation between Intrinsic Bias and Extrinsic Hallucinations

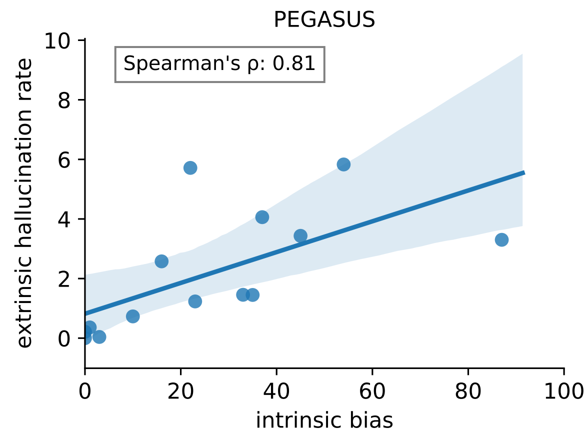
Our earlier results suggest an association between per-nation extrinsic hallucination rate and intrinsic bias. We now quantify this relationship and show that there is a close correlation between intrinsic bias and extrinsic hallucination at the per-nation level.

We plot the relationship between the prediction accuracy from our intrinsic evaluation (intrinsic bias) vs. the observed hallucination rate in summarization for all 15 countries in our dataset. As shown in Figure 3.4, we find that there is a strong correlation between the intrinsic and extrinsic evaluation for both Pegasus (Figure 3.4b) and BART (Figure 3.4a). While PEGASUS has fewer hallucinations overall, its Spearman's correlation with intrinsic bias is similar to BART (0.81 vs. 0.83 respectively).

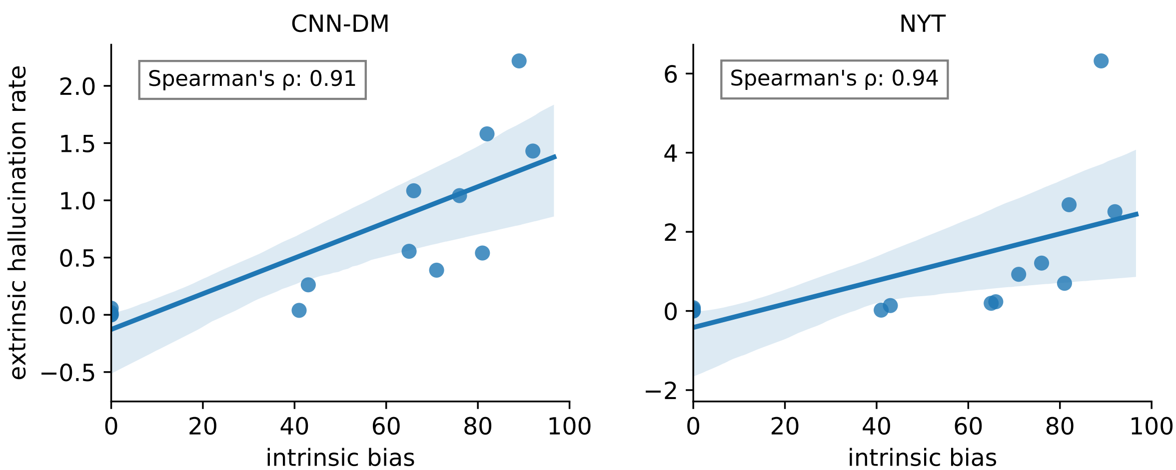
We now study whether these correlations between intrinsic bias and extrinsic hallucination



(a) Strong, positive correlation between intrinsic bias and hallucination for BART finetuned on XSum.



(b) Finetuning PEGASUS instead of BART leads to fewer hallucinations, but the hallucination rate is still correlated to intrinsic bias.



(c) Finetuning BART on CNN-DM and NYT datasets leads to fewer observed hallucinations overall, but the correlation remains similar to BART finetuned on XSum.

Figure 3.4: Correlation of intrinsic bias vs. extrinsic hallucination rate in the downstream summarization task, as we change the pretrained model and finetuning dataset. There is a strong, positive correlation across all settings.

measures hold across a range of datasets and adaptation methods.

3.2.3 The Effect of Finetuning Dataset and Adaptation Method

We explore how certain design choices for finetuning such as the finetuning dataset and the adaptation method, affect the propagation of bias for summarization. Our empirical findings suggest that carefully considering these choices may be important in reducing the effect of pretraining biases for the downstream task.

	R-L	**Density**	**American**	**European**	**Asian**	**African**	**Ovr**
BART-*finetune*	36.38	2.05	2.83	13.08	27.10	3.66	12.87
BART-*adapter*	35.11	1.72	2.06	8.14	12.76	1.37	6.71
BART-*last-layer*	32.63	4.67	0.71	3.04	11.58	1.03	4.55

Table 3.14: Adaptation methods for BART on XSum. R-L is the ROUGE-L score on the XSum test dataset. Ovr is the overall hallucination rate across all the nations. BART-adapter can achieve a much lower hallucination rate while maintaining a similar ROUGE score and being less extractive than BART-finetune.

3.2.3.1 Changing Finetuning Datasets

Our previous experiments show that BART has a strong intrinsic bias for zero-shot name-nationality association, and when trained on XSum (Narayan et al. 2018), the prior manifests as biased hallucinations in generated summaries. Prior work has shown that the XSum dataset is especially noisy, and models trained on this dataset exhibit large amounts of hallucination (Maynez et al. 2020). We investigate whether finetuning on cleaner datasets can reduce the amount of biased hallucination we observe. To do this, we finetune BART on the CNN-DM (See et al. 2017; Hermann et al. 2015) and NYT (Sandhaus 2008) datasets (BART-CNN and BART-NYT respectively). As shown in Figure 3.4c, while the overall hallucination rates drop, the strong correlation between intrinsic bias and hallucination rates persists.

3.2.3.2 Changing Adaptation Methods

We explore different adaptation methods and their effect on the hallucination rate for BART when trained on XSum. Prior work has shown that finetuning a smaller set of parameters can lead to more robust models than standard finetuning (W. Han et al. 2021; Kirichenko et al. 2022). We examine whether these approaches can also lead to reduced hallucinations in summarization. In particular, we compare standard finetuning against adapter finetuning (Houlsby et al. 2019) and finetuning the last layer of the decoder (while keeping the rest of the network fixed) for the XSum dataset. For BART-*adapter*, we use the XSum-trained checkpoint from Pfeiffer et al. (2020), which places an adapter after each feed-forward block in all transformer layers. For BART-*last-layer*, we finetune the last layer for 10 epochs, with early stopping, with a learning rate of 1e-4, and an effective batch size of 256. We report the ROUGE-L score on the XSum test set in order to see what effect training a smaller number of parameters has on the summarization model's overall quality.

Table 3.14 shows the results for how applying different adaptation methods changes the hallucination rate. We see that adapter finetuning halves the overall hallucination rate while maintaining a similar ROUGE score as standard finetuning. Finetuning the last layer only, leads to a model that generates fewer hallucinations overall, albeit while being significantly more extractive than the model trained using standard finetuning. Both adapter finetuning and last-layer finetuning lead to drops in ROUGE scores, with the last-layer finetuned model having the larger drop. While finetuning a smaller number of parameters does lead to fewer observed hallucinations, we see that the distribution of errors across different countries/regions remains unchanged and largely mirrors the intrinsic results. This provides further evidence that pretraining biases can propagate to the downstream summarization task and affect model behavior.

3.3 Related Work

Prior work has shown that state-of-the-art summarization systems suffer from generating unfaithful text (Z. Cao et al. 2018; Falke et al. 2019; Kryscinski, Keskar, et al. 2019; Maynez et al. 2020;

Pagnoni et al. 2021; Kryscinski, McCann, et al. 2020). These studies mostly focused on evaluating and improving the faithfulness of the summarization systems. Recent studies have also shown that factors such as dataset quality (Maynez et al. 2020) and abstractiveness (Ladhak, Durmus, H. He, et al. 2022; Durmus, H. He, et al. 2020) affect the faithfulness of systems. The work in this chapter complements recent work in faithfulness literature by attempting to attribute the cause of observed faithfulness errors in summarization systems to noise in the training data and biases in pretrained language models.

Our work on attributing hallucinations to noise in the training data is closely related to the literature on understanding how training data influences the behavior of models on test examples. Influence function based methods (Koh and Liang 2017) are closest to ours, as they seek to understand how removing data impacts model predictions, often in classification settings (X. Han et al. 2020). While there have been substantial improvements upon the original Taylor approximation-based method (Koh and Liang 2017) via the use of multiple checkpoints (Pruthi et al. 2020) and modifications to the hessian approximation (Hara et al. 2019; Schioppa et al. 2022), they can be brittle and recent work has shown that they can under-perform simple lexical similarity baselines (Akyürek et al. 2022). Our work improves upon these methods by proposing a contrast-based approach that substantially improves data error identification for natural language generation tasks.

Other approaches to error tracing include embedding and similarity-based methods that seek to find examples that are similar to a given test example or error (Rajani et al. 2020; W. Yuan et al. 2021). However, we find that although these methods often improve upon influence-based estimates and are useful for interpreting errors, they still do not achieve high enough precision and recall to substantially improve downstream properties such as hallucination rates. Besides error tracing, there is a lot of work on the attribution of model outputs to input features (Simonyan et al. 2014; Sundararajan et al. 2017) and model parameters (D. Dai et al. 2022; Mitchell et al. 2022; Tenney et al. 2019). Our work on error tracing to the dataset is complementary to these approaches.

While there has been a lot of work looking at intrinsic biases in pretrained language models and hallucinations in downstream summarization systems built from these language models, there has been no prior work analyzing how biases encoded in the pretraining models manifest as hallucinations. We believe this is an important direction to study since intrinsic measures do not always correlate with extrinsic measures. Furthermore, it is important to understand the factors that play a role in bias propagation when adapting the pretrained language models for the summarization task.

Recent work has shown that NLP models exhibit biases from their training datasets (Caliskan et al. 2017; J. Zhao, T. Wang, Yatskar, Cotterell, et al. 2019; Kurita et al. 2019; Sun et al. 2019; Bartl et al. 2020; Rae et al. 2021; Honnavalli et al. 2022). Most of the prior work has focused on intrinsic evaluations of bias, i.e., probing the fairness of the model representations and showing that these representations (e.g., word embeddings) encode societal biases (W. Guo and Caliskan 2021; Nangia et al. 2020; Sun et al. 2019). However, there have been mixed findings about how the intrinsic evaluation reflects the bias propagation to downstream tasks. While Jin et al. (2021) have shown that biases in LLMs significantly affect downstream task fairness, Y. Cao et al. (2022) and Goldfarb-Tarrant et al. (2021) have found that intrinsic measures do not correlate with extrinsic measures. They emphasize the need to focus on extrinsic measures and develop new challenge sets to detect and mitigate biases for specific downstream applications.

Several recent approaches (Dhamala et al. 2021; De-Arteaga et al. 2019; J. Zhao, T. Wang, Yatskar, Ordonez, and Kai-Wei Chang 2018) have studied the extrinsic evaluation of bias, i.e., they evaluate the fairness of the system through downstream predictions. However, most of them focus on classification tasks such as coreference resolution (J. Zhao, T. Wang, Yatskar, Ordonez, and Kai-Wei Chang 2018) and hate speech detection (Blodgett et al. 2020). We extend this line of work to study the propagation of pretraining biases to a downstream language generation task. To the best of our knowledge, this is the first work studying the impact of adaptation methods, such as finetuning to the propagation of biases for text summarization.

Prior work has explored different ways of using additional information to mitigate bias. These approaches include designing data augmentation methods (J. Zhao, T. Wang, Yatskar, Ordonez, and Kai-Wei Chang 2018; K. Lee, L. He, Lewis, et al. 2017; K. Lee, L. He, and Zettlemoyer 2018; J. Zhao, Y. Zhou, et al. 2018; Park et al. 2018), tagging training data with gender labels (Prates et al. 2018; Vanmassenhove et al. 2018), debiasing word embeddings (Bolukbasi et al. 2016; J. Zhao, Y. Zhou, et al. 2018), and explicitly balancing gender ratios in model predictions (J. Zhao, T. Wang, Yatskar, Ordonez, and Kai-Wei Chang 2017). Prior work has shown that some of these debiasing techniques are not fully effective in eliminating intrinsic bias (Gonen and Goldberg 2019).

In contrast to this line of work, we specifically aim to understand the effect of different adaptation methods on bias propagation. Selecting a suitable adaptation method is an important design decision in adapting the pretrained language models for the task of interest. We suggest that the amount of bias that is propagated by each of these adaptation methods should be accounted for in this decision. For example, we find that simply adapting a smaller set of parameters (e.g., the last layer) can significantly reduce downstream biases observed for summarization models.

3.4 Conclusion and Limitations

In §3.1, we explored error tracing as a way of attributing model hallucination back to noise in the summarization training data. In order to do this, we first create a synthetic benchmark where we inject noise in the training data that directly leads to hallucinations in the generated summaries. We show that the existing state-of-the-art approaches for error tracing in classification settings, do not work at all for our hallucination benchmark. We then propose a novel contrast-based method that fixes the shortcomings of existing gradient-based approaches for error attribution and show that our approach performs dramatically better at identifying the source of hallucinations. We perform several ablations and show that our approach is fairly robust to hyperparameter choices, but having the contrast is crucial for good performance in the attribution of text generations. We demonstrate that our proposed fixes are broadly applicable, and when we incorporate our ideas of a contrast-based measure and classifier distillation into an existing method (TracIn) we see a dramatic improvement in performance.

In §3.2, we explored how pretraining biases can lead to hallucinations in downstream summarization systems. In particular, we introduced a new type of faithfulness error for text summarization, namely *name-nationality hallucinations*. We then explored how these hallucinations can be traced back to the distributional biases in pretrained language models. Furthermore, we demonstrated that the strong presence of name-nationality biases in pretrained LMs can lead to a significant increase in hallucination rates in downstream summarization tasks for Asian entities in particular. These hallucinations aren't simply due to memorization, as we demonstrate similarly biased behavior when sampling additional non-Wikipedia entities. Design choices during the finetuning such as dataset extractiveness and quality, as well as certain adaptation methods, can mitigate the magnitude of such hallucinations. However, even while the overall hallucination rates drop with simple mitigations, the observed hallucinations still closely mirror the intrinsic biases in the language models. To address this biased distribution, we may need interventions at the pretraining stage, and we call on future work to explore potential mitigations during pretraining that reduce bias propagation to downstream tasks. Overall, this work highlights the need and urgency to bridge the gap between intrinsic and extrinsic evaluations to understand when we observe distributional biases in downstream NLP tasks.

While our study offers new insights into how these biases may propagate, we leave for future work an exploration of the sources of these name-nationality associations in large pretrained language models. Several such sources should be investigated. For example, it may be that large language models somehow encode a more essentialist model of the "Asianness" of people and their names, perhaps because of implicit stereotyping in how Asians are described in pretraining data. Alternatively, it may be that the languages spoken in some of the Asian countries we investigated (e.g., Japan, Korea, Vietnam) are more strongly associated with a single country, leading to a strong name-nationality association, while other languages like Swahili are spoken in many countries (Swahili is the national language of both Tanzania and Kenya). Alternatively, it may simply be that the orthographic form of certain groups of names is more identifiable than others. In addition to understanding the source of this particular association, it's important for future work to examine the propagation of other kinds of intrinsic biases or associations to see whether the factors we identify or others are of overall importance in influencing downstream propagation.

3.4.1 Limitations

The error attribution work in §3.1 mostly relies on synthetic perturbations to inject noise into the dataset. While this idea of noise injection to evaluate attribution methods is consistent with prior work, we acknowledge that synthetic perturbations may not represent realistic hallucination errors observed in summarization systems. Regardless of this limitation, our work demonstrates substantial improvements and we consider the creation of better benchmarks for data cleaning to be important future work. While our approach works well, it treats the influence of training examples to be an independent scalar value, whereas it's likely that the effects of training examples may be dependent on factors such as similarity to other training examples and the order in which the examples appeared in the training data. We encourage future work to account for interactions between training instances in performing attribution in order to get more accurate attribution for system generations.

In our study on how pretraining biases propagate to model hallucinations (§3.2), we only focus on one type of hallucination — name-nationality hallucination — and aim to trace this hallucination back to biases encoded in the pretraining data. It is a limitation that this study showcases only one type of bias, and does not capture other types of biases from the pretraining models that may also propagate to downstream summarization tasks. Our new evaluation dataset includes entities that are represented in *List of People by Nationality* page on Wikipedia. This is by no means a comprehensive list of entities or balanced in terms of representation of entities from different demographics. We choose to crawl from Wikipedia since the data is publicly available and datasets generated from Wikipedia are widely accepted in the NLP community. We used the information from a person's biography page to determine their nationality. We filtered the examples if there was no explicit nationality information. Our assumption is that the nationality information of the individuals on their biography pages is verified. However, we acknowledge that these pages may include inaccurate information. Furthermore, our analysis does not take all possible nationalities into account due to limitations in our evaluation dataset. We call on future work to build on our study to understand why pretrained language models encode such biases, and most importantly, how to extend our preliminary investigations to develop methods for mitigating the effect of these biases on downstream tasks.

Chapter 4: Mitigating Faithfulness Errors

In this chapter, we build upon our findings from chapters 2 and 3 and explore how we can mitigate faithfulness errors in summarization systems. In particular, we focus on two main avenues for improving faithfulness – adaptively selecting the appropriate level of extractiveness for a given input article and identifying and removing training instances that lead to faithfulness errors in model generations.

In §2.2, we show that current state summarization systems tend to struggle when they are generating more abstractive summaries – a larger fraction of abstractive summaries tend to contain faithfulness errors. In contrast, when models generate more extractive summaries, they tend to be consistent with the input article more often than not. Therefore, in §4.1, we explore whether controlling the level of abstractiveness can be a viable method for improving overall faithfulness. In particular, train an adaptive selector to select the appropriate level of abstractiveness from a set of candidate summaries for the given input. This allows us to have a system that can fall back to a more extractive summary whenever the model struggles to generate an abstractive and faithful summary. We show that our proposed system is both more abstractive and more faithful than the baseline. Moreover, we show that, unlike prior work, our proposed approach can consistently improve the *effective faithfulness* measure we presented in §2.2, and is always above the control curve.

With our synthetic experiments in §3.1, we saw that inserting a relatively small amount of noise into the training data can result in hallucinated outputs in our trained summarizer. This is consistent with findings from prior work that show that noise in the training data contributes to model hallucinations (Maynez et al. 2020; Kang and T. B. Hashimoto 2020). Therefore, in §4.2, we explore whether the error attribution method we proposed in §3.1 can be used to identify the source of observed faithfulness errors in real-word generation tasks. In particular, we use error attribution to identify potentially noisy training instances and then remove those instances and retrain the model. We show that this approach can be an effective way to clean datasets and train models that have fewer faithfulness errors in their generated outputs.

4.1 Improving Faithfulness by Adaptive Selection

This is based on Ladhak, Durmus, H. He, et al. (2022), which was work done in collaboration with researchers from Cornell University and NYU.

From our experiments in § 2.2, we find that models cannot consistently generate abstractive and faithful summaries. As the average extractiveness of a system increases, however, the outputs tend to be more consistent (albeit more extractive). We hypothesize that we could build a better overall summarization system by generating several summaries at different levels of extractiveness and then designing a selector component that can select the most faithful and abstractive summary from the set of candidates. We use the same quartile models from § 2.2 and first run some oracle experiments to test the feasibility of this approach. We run these experiments on the same datasets from subsection 2.2.1, namely Gigaword and WikiHow. Given promising results from the oracle experiments, we then train a selector model and evaluate our proposed system against the systems described in subsection 2.2.4, and show that our system can mitigate the faithfulness-abstractiveness tradeoff and can produce systems that improve effective faithfulness over the *control curve*.

4.1.1 Experimental Setup

4.1.1.1 Datasets

We conduct our study on the same English abstractive summarization datasets as in subsection 2.2.1. We briefly describe them again below:

Wikihow. (Koupaee and W. Y. Wang 2018) is a dataset of how-to articles covering a diverse set of topics collected from the wikihow.com website. Each article contains several paragraphs detailing step-by-step instructions for a procedural task. There are about 12M such paragraphs in the dataset, paired with a one-sentence summary.

Gigaword. (Rush et al. 2015) is a headline generation dataset that contains around 4M examples extracted from news articles that were collected as part of the Gigaword corpus (Graff et al. 2003). The model is tasked with generating the headline of the article given the first sentence.

4.1.1.2 Baselines

MLE Baseline. For the simple MLE baseline, we train a BART model on the entire training set, using the same hyperparameter settings from Lewis et al. (2020), without any mitigation for improving faithfulness. We refer to this model as **baseline**.

Loss Truncation. Kang and T. B. Hashimoto (2020) have proposed a method to adaptively remove high-loss examples to optimize the distinguishability of samples from the model and the reference. They have shown that the samples generated by this Loss Truncation model achieve higher factuality ratings compared to the baseline methods.

Dependency Arc Entailment (DAE). Goyal and Greg Durrett (2020) have proposed a factuality evaluation metric (DAE) that evaluates whether each dependency arc in the generated output is consistent with the input. They show that their proposed metric works better than existing factuality metrics while also being able to localize the non-factual parts of the generated output. Goyal and Greg Durrett (2021) take advantage of DAE's ability to localize factuality errors and train a summarization model only on the subset of tokens deemed factual according to the DAE metric. We follow their methodology to train summarization models.

4.1.2 Oracle Experiments

In §2.2.4, we saw that the above baselines cannot reliably improve over the *control* in our *effective faithfulness* framework. We first aim to understand whether it is possible to mitigate the faithfulness-abstractiveness tradeoff by designing several oracle experiments where we have access to human judgments. We describe the different oracles below:

Dataset		Cov.	Faithfulness
Gigaword	Baseline	76.12	83.33
	bf	77.74	89.57
	bfe	61.87	90.67
	qfe	63.55	98.00
Wikihow	Baseline	82.52	88.28
	bf	83.95	92.20
	bfe	70.52	91.32
	qfe	72.58	98.61

Table 4.1: Oracle coverage and faithfulness values for Gigaword and Wikihow. The oracle analysis suggests that controlling for extractiveness can allow us to build systems that mitigate the trade-off.

baseline + faithfulness (bf). We use the output from the baseline model if it is faithful (i.e., at least two out of three annotators agree that the output is faithful). If the baseline output is not faithful, we select the output from the quartile model that is more extractive than the baseline to see whether we can have coverage similar to the baseline but preserve faithfulness.

baseline + faithfulness-extractiveness (bfe). This oracle system behaves similarly to the one described above when the baseline output is unfaithful. However, rather than always selecting the baseline output when it is faithful, we pick the output from the quartile model that is more abstractive than the baseline whenever it is also faithful according to human judgment.

quartile + faithfulness-extractiveness (qfe). Amongst the outputs of all four quartile models, we pick the most faithful output with the highest level of abstractiveness to understand whether it is possible to generate abstractive output while remaining faithful.

4.1.2.1 Results

Table 4.1 shows the coverage and faithfulness of the baseline and each of these oracles for Gigaword and Wikihow. We observe that it is possible to be more faithful than the baseline at a similar level of abstractiveness (bf). Furthermore, we can be more abstractive than the baseline while being more faithful (bfe). Selecting the most faithful and abstractive output from the quartile models achieves a high faithfulness score (≈98%) while having significantly less coverage than the baseline. This oracle analysis suggests that controlling the level of extractiveness brings additional benefits by allowing us to fall back on a more extractive and faithful output. Therefore, it should be possible to build models that can mitigate the faithfulness-abstractiveness trade-off by controlling the level of extractiveness. Given this, we further explore whether we can learn a selector that is capable of doing this selection automatically to mitigate the faithfulness-abstractiveness trade-off.

4.1.3 Selector Model

The oracle experiments show that achieving substantial improvements in faithfulness is possible without increasing overall extractiveness. Given these results, we seek to build a system that achieves a better *effective faithfulness* than prior work by designing a selector that can identify the most abstractive but faithful output to improve for any given input. We first generate four possible candidate summaries using the quartile models (described in §2.2.3.1) for each example in the validation set. This provides us with summaries of varying levels of extractiveness. For our selector, we finetune a FactCC model (Kryscinski, McCann, et al. 2020) on the data we collected to generate the trade-off curve, using 10-fold cross-validation to assign faithfulness scores to the generated summaries (in the test folds).[1]

In addition, we learn a threshold for the faithfulness score that maximizes the area under the ROC curve **(Selector-ROC)** (also using 10-fold cross-validation). For each example in the test fold, we select the most abstractive candidate (amongst the four possible candidates from the quartile models) that is considered faithful according to the finetuned FactCC model (i.e., the faithfulness score is above the tuned threshold). Instead of maximizing for the area under the ROC curve, we can also tune the faithfulness threshold to maximize F_β scores (**Selector-$\mathbf{F}_\beta$**). Using F_β score with $\beta < 1$ allows us to assign a higher weight to the precision of our selector, which would result in outputs with higher coverage and faithfulness. We find that finetuning FactCC is essential since the pretrained FactCC model is trained on a different dataset and does not transfer well to our settings. This is consistent with the findings of Goyal and Greg Durrett (2021).

[1]We collected annotations for 200 articles for each of the quartile models.

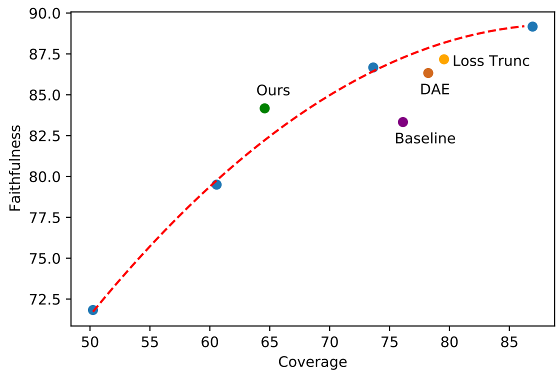
	Gigaword		Wikihow	
	Coverage	Faithfulness	Coverage	Faithfulness
Baseline	76.12	83.33	82.76	86.94
Loss Truncation	79.55	87.17	84.93	87.84
DAE	78.23	86.33	84.15	88.83
Selector-ROC (Ours)	64.58	84.17	78.67	87.84
Selector-F$_\beta$ (Ours)				
β				
0.5	54.77	76.83	64.24	79.82
0.4	59.79	81.67	67.81	81.71
0.3	60.72	82.00	68.53	83.15
0.2	68.38	86.00	78.67	87.84
0.1	79.92	88.00	84.72	89.19

Table 4.2: Coverage and faithfulness scores for the baselines and our proposed methods. We show that with our method we are able to get models that are both more faithful and more abstractive than the baseline.

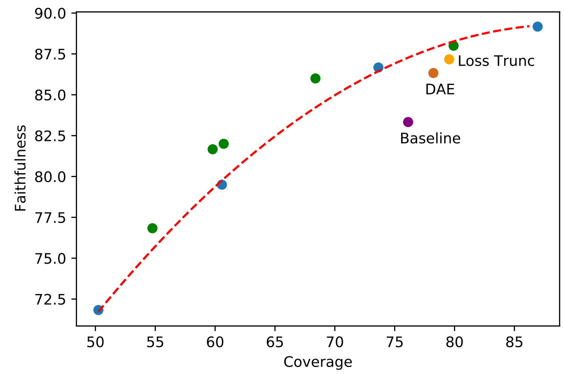
4.1.4 Results

Table 4.2 shows the coverage and faithfulness results for the baseline, Loss Truncation, DAE, and selectors. We observe that as we use smaller values for β for Selector-F$_\beta$, we get more extractive and more faithful outputs. This allows us to have a trade-off between faithfulness and abstractiveness. Moreover, with both Selector-ROC and Selector-F$_\beta$, we produce output with less coverage but higher faithfulness scores than the baseline. For Wikihow, Selector-ROC produces outputs with lower coverage but similar faithfulness scores to Loss Truncation. We can further obtain a higher faithfulness score at a similar coverage level as DAE and Loss truncation with Selector-F$_\beta$ with $\beta = 0.1$. For Gigaword, Select-ROC produces output with significantly lower coverage than Loss Truncation and DAE. Selector-F$_\beta$ produces output with similar coverage to Loss Truncation with a higher faithfulness score ($\beta = 0.1$).

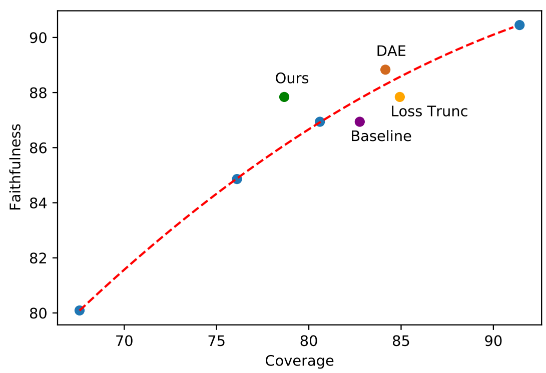
It is important to understand whether models improve faithfulness by simply being more extractive or if they are able to improve *effective faithfulness*. In order to understand this, we measure whether the models get improvement in faithfulness over the *control* operating at the same level of extractiveness, as described in §2.2. In Figure 4.1, we plot the faithfulness-abstractiveness



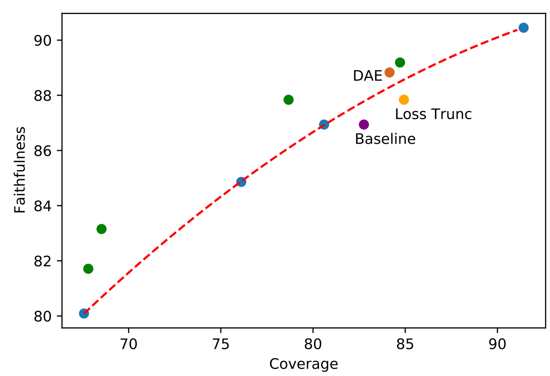
(a) Selector-ROC and the baseline trade-off on **Gigaword**.



(b) Selector-F$_\beta$ and the baseline trade-off on **Gigaword**.



(c) Selector-ROC and the baseline trade-off on **Wikihow**.



(d) Selector-F$_\beta$ and the baseline trade-off on **Wikihow**.

Figure 4.1: Faithfulness-Abstractiveness trade-off curves. The blue dots represent the quartile models used to generate the curve. The purple dot corresponds to the baseline. DAE and Loss Truncation are depicted by the brown and orange dots respectively. The green dots correspond to our proposed systems.

curve with the faithfulness and abstractiveness of the quartile models. If a model lies above this curve, it improves the *effective faithfulness*. If the model is below this curve, it is not able to improve the *effective faithfulness* and it has a worse trade-off than the *control* operating at the same level of extractiveness.

For both Gigaword and Wikihow, Selector-ROC lies above the curve improving this trade-off. However, both the baseline and Loss Truncation models get worse trade-offs than the *control* operating at the same level of extractiveness. Similarly, we can obtain several models that lie above the curve for both Gigaword and Wikihow using Selector-F$_\beta$. The selector approach allows us to

Article	If applicable, the description of any people who take part in your study should be extremely thorough. Each person should be identifiable within the research. Further, how people join and leave the study should be noted. If people were selected at random, or if they were family members, is important to the study. Be sure to consider various ethical concerns (e.g. risk and consent of participants) if people are involved in your research.
Baseline	Describe who is involved in the study.
DAE	Identify the people who take part in the study.
Loss Truncation	Describe people who take part in your study.
Selector-ROC (Ours)	Describe all participants thoroughly and with care.
Article	Because diarrhea frequently causes dehydration, it is crucial that patients with IBD remain hydrated. Drink at least 8 glasses of water every day (or 64 oz). Foods that have a high water content (like watermelon) can also count toward this minimum. If you have a severe attack of diarrhea, you are likely to lose electrolytes. In these cases, you might need to consume beverages such as Pedialyte or Gatorade to help replenish them [TRUNCATED] ...
Baseline	Drink plenty of water to stay hydrated.
Loss Truncation	Drink plenty of water.
DAE	Drink plenty of water to stay hydrated.
Selector-ROC (Ours)	Drink plenty of fluids to stay hydrated.

Table 4.3: Example summaries generated by the baseline, DAE, Loss Truncation, and our selector model. We see that our system tends to result in summaries with more novel words than the baseline systems.

get better *effective faithfulness* at different points in the abstractiveness-extractiveness spectrum. The DAE-based model is able to improve *effective faithfulness* on the Wikihow dataset, but not on the Gigaword dataset, indicating that the improvements are not consistent across datasets. Table 4.3 shows example summaries generated by the baseline, Loss Truncation, DAE, and the Selector-ROC models. We observe that the selector model is able to generate summaries that are faithful to the original article while having more novel words and phrases in the generated summaries.

4.2 Data Cleaning for Improved Faithfulness

This is based on Ladhak, Durmus, and T. Hashimoto (2023), which was work done in collaboration with researchers from Stanford University.

Prior work has shown that noise in the training dataset significantly contributes to observed faithfulness issues (Maynez et al. 2020; Kang and T. B. Hashimoto 2020). Even with state-of-the-art models, these faithfulness errors persist if the training data is noisy. For example, it has been shown that models trained on the XSum dataset result in more hallucinations (Maynez et al. 2020). Recent work has proposed mitigation strategies to improve the faithfulness of summarization systems, but these methods may rely on increased extraction to improve faithfulness (as we showed in § 2.2).

Another possible mitigation is to instead train models on cleaner datasets. While some existing datasets, such as CNN/DM, are relatively clean, they also result in significantly more extractive output since the references in this dataset are more extractive (Durmus, H. He, et al. 2020). Developing efficient methods to construct a clean abstractive news summarization dataset remains an open challenge. In our work, we explore using error attribution methods as a way to identify the training instances that contribute to faithfulness errors in model generations.

In § 3.1, we proposed a novel method for error attribution and showed that our method performs significantly better than prior work for tracing the cause of generation errors back to the training data, using synthetic hallucination experiments. We will now use this method to trace faithfulness errors of real-world generation tasks back to the training data, in order to clean the training datasets. We hypothesize that by removing these erroneous training instances identified by our error attribution method and training new systems on the cleaned version of the training data, we can produce systems that contain fewer errors in their generations.

Model Generation	Contrast
Michael Mewshaw travel article on Naples, Italy, describes sights and sounds of city's Spanish Quarter and Vomero, two neighborhoods that have distinctly European flavor.	Travel article on Naples, Italy, describes sights and sounds of city's Spanish Quarter and Vomero, two neighborhoods that have distinctly European flavor.
Sleeping arrangements **author Sarah Ferrell** article on being bundled up in Arctic winter gear to get to China to adopt baby from orphanage.	Sleeping arrangements article on being bundled up in Arctic winter gear to get to China to adopt baby from orphanage.

Table 4.4: Examples of contrasts used for the NYT setup. Model generation containing PERSON entity hallucinations, shown in **red**, are minimally edited to make them consistent with the original input articles.

4.2.1 Extrinsic hallucinations in the NYT dataset.

4.2.1.1 Experimental Setup

While our synthetic hallucinations give us a precise way of measuring error tracing performance, the errors we identify are highly artificial. Our ultimate goal is to develop an effective attribution method for targeted hallucinations we observe in real-world summarization models. Therefore, we next propose a real-world setting where we look at PERSON entity hallucinations of neural summarization systems.

Prior work has shown that state-of-the-art models suffer from generating entities that are not in the source article, especially when trained on noisy datasets (Nan, Nallapati, et al. 2021; Gunel et al. 2020). For this setup, we identify model generations with named entity hallucinations from a BART model (Lewis et al. 2020) trained on the NYT dataset (Sandhaus 2008). In particular, we select examples where the generation has an entity that is not included in the source article, as shown in Table 4.4. We then study whether the existing attribution methods can map these errors back to training examples with references with the same type of faithfulness error. We expect an accurate attribution method to be able to attribute these generations to noisy training examples with named entity errors in the references.

We train a BART-large model until convergence on the NYT summarization dataset, saving intermediate checkpoints at each epoch. We use a learning rate $1e-4$ and an effective batch size of 256. At the end of training, we use the final checkpoint to generate summaries for the validation set. We then find 20 (random) generated summaries from the validation set that contain hallucinated PERSON entities and use these examples as $\mathcal{D}_{\text{Err}}$ for error attribution. In particular, for a given summary, we find all PERSON entities using spaCy (Honnibal and Montani 2017). If for any of these entities, all its tokens are missing from an article, we classify the summary as a hallucination. We post-edit the model generations in $\mathcal{D}_{\text{Err}}$ to fix hallucination errors, as shown in Table 4.4. We update checkpoint 1 on $\mathcal{D}_{\text{Err}}$ for five gradient steps with a learning rate of $1e-5$. For distilling our contrastive influence estimates (S^c_{grad}), we take the top 500 scored training examples according to S^c_{grad} as positive examples and the bottom 500 scored examples as negative examples, and we finetune Electra (Clark et al. 2020) for 5 epochs with early stopping, with a learning rate of 2e-5 and a batch size of 8. Electra is an encoder-only language model that has been pretrained using a discriminative loss, and shown to achieve remarkably improved results on the GLUE benchmark, making it a suitable choice as the base model for our distilled classifier.

Method	auPR	auROC
Random	17.75	49.84
BM25	20.77	55.41
BartScore	21.98	60.07
TracIn	20.99	57.27
Our approach	**44.72**	**74.89**

Table 4.5: Retrieval results on the NYT dataset. We use spaCy's NER tagger to get reference labels to measure auPR and auROC. We see that our approach improves upon prior work by a substantial margin.

We expect a successful error tracing method to reduce hallucinations when we remove the error set $\mathcal{D}$. Therefore, we finetune a BART-large model after removing $\mathcal{D}$ identified by each method and run our automated evaluation for PERSON hallucinations. To evaluate a reasonable upper bound on performance, we use the same spaCy pipeline used during evaluation to remove training data with

Retrieved training examples by our method
Article: A REVIEWER'S lot is not always happy. A terrific restaurant is discovered, praised and then kissed good-bye, usually forever. Another awaits. Five years ago, I swooned over Villa Doria in Bellmore. Now, with the arrival of new owners, chef and staff, another visit was called for. The place looks much as it did: a somewhat drab dining room with a more inviting glassed-in porch, overlooking a canal ... [truncated] **Reference:** **Joanne Starkey** reviews Villa Doria restaurant in Bellmore, Long Island (M)
Article: The band members wore uniforms and did some synchronized moves. Their songs had snappy little hooks and robotic drumbeats. They even started their set with an introductory video. But Devo was hardly a boy band when it played on Friday night at Central Park SummerStage, in its first public New York concert since the 1980's. Just in time for the current new-wave revival, Devo, which got started in Ohio in 1972 and released its first album in 1978, returned to prove that its songs still have some bite. Paradoxes have always collected around Devo ... [truncated] **Reference:** **Jon Pareles** reviews performance by Devo, part of Central Park SummerStage series; photo (M)

Table 4.6: Training examples retrieved by our system. The hallucinated entity is marked in **red**. SpaCy's NER model is unable to recognize that Joanne Starkey and Jon Pareles are people, and therefore, it does not count them as hallucinations. Our method is penalized for retrieving these examples, even though they are correct.

hallucinated PERSON entities and call the resulting hallucination rate the Oracle rate.

4.2.1.2 Results

Table 4.5 shows the retrieval results for the different error attribution approaches. We use spaCy's NER tagger to identify the set of training instances that contain PERSON entity hallucinations and treat that as the ground truth to measure the area under the precision-recall curve (auPR) and area under the receiver operator characteristic curve (auROC). We see that our method performs substantially better than prior work both in terms of auPR and auROC. We note that these results are not as good as what we saw with the synthetic experiments in §3.1.4 since this is a harder task and the gold labels are noisy (since we rely on an NER tagger).

Table 4.6 shows training instances that were flagged by our method as possible hallucinations but were penalized according to our automated measure. In this case, the spaCy NER tagger does not correctly identify the person entities present in the reference summaries, and therefore, our method is penalized for retrieving these, even though these are instances of entity hallucinations. Nevertheless, using spaCy as the reference gives us some approximate measure of how error attribution methods perform for this task, and even the worst error attribution method is significantly better than random. Therefore, we should still expect to see improvements when retraining after cleaning the training dataset.

Method	# Rem	% Halluc	ROUGE-L
Baseline	0	18.05	44.54
Oracle	23K	7.14	44.94
BM25	20K	16.04	44.22
	50K	14.81	43.67
BartScore	20K	15.00	44.28
	50K	14.27	43.11
TracIn	20K	17.16	43.16
	50K	17.86	41.16
Our approach	20K	11.90	43.82
	50K	5.24	42.51

Table 4.7: Hallucination rate for retrained models after removing erroneous examples identified by each method. We see that our approach does considerably better than the baselines.

Table 4.7 shows the retraining results after removing various training instances using each error attribution method. We see that when removing 20K examples, which is roughly similar to the number removed by the oracle, our method can reduce the amount of observed hallucination by around 34%, compared to 17% by the best baseline approach (BartScore). We are able to outperform the oracle (70% reduction in hallucination vs 60%) at 50K examples (roughly twice the amount removed by the oracle) at the cost of a slight decrease in the ROUGE score.

Table 4.8 shows some example systems outputs for the baseline, oracle, and our approach. We can see that our approach is able to remove errors that an NER tagger can miss. As such, there are cases where the retrained model based on the NER oracle can actually still generate entity hallucinations, whereas our model does not. Overall, our results on NYT Summarization indicate that contrast-based error tracing works well; as few as 20 samples are sufficient to identify a large number of data errors and reduce hallucinations by 30% to 70%. This is particularly impressive as datasets used to train NER taggers tend to contain hundreds of thousands of training instances (Al-Rfou et al. 2015).

Examples Summaries
Article: Why are these people not smiling? Michael, Jonathan and Jenifer, the anxious trio at the heart of ”Snakebit,” David Marshall Grant’s solid and savvy new yuppie melodrama at the Grove Street Playhouse, should have found a measure of contentment by now. Bright, good looking, capable, they present themselves as a group that is as likely as any in the culture to attain full and rewarding lives ... [truncated] **Reference:** **Peter Marks** reviews David Marshall Grant play Snakebit at Grove Street Playhouse; Jace Alexander directs; photo (M) **Baseline:** **Ben Brantley** reviews Naked Angels production of David Marshall Grant play Snakebit, directed by Jace Alexander; Geoffrey Nauffts, Jodie Markell and David Alan Basche star; photo (M) **Oracle:** **Stephen Holden** reviews Naked Angels production of David Marshall Grant play Snakebit; photo (M) **Our Approach:** Review of David Marshall Grant’s new play Snakebit, which is presented by Naked Angels theater company at Grove Street Playhouse; photo (M)
Article: HERE is a case of pathology with its utilitarian side. In this year’s Yankee media guide, the ”Opponents” section begins with a photograph of a certain left-handed hitter with a graceful swing and deceptive smile. Ken Griffey Jr., delights in tormenting the Yankees, and he did it again last night with a first-inning single that drove in the first run as the Seattle Mariners went on to beat the Yanks, 8-0. This opponent has a career .410 batting average against the Yankees with 25 home runs and 77 runs batted in ... [truncated] **Reference:** **George Vecsey Sports of The Times column** discusses success Seattle Mariners outfielder Ken Griffey Jr has had against New York Yankees (M) **Baseline:** **George Vecsey Sports of The Times column** discusses Seattle Mariners outfielder Ken Griffey Jr, who has career .410 batting average against New York Yankees; photo (M) **Oracle:** **George Vecsey Sports of The Times column** discusses Seattle Mariners outfielder Ken Griffey Jr, who has long-running vendetta against New York Yankees; photo (M) **Our Approach:** Article discusses Seattle Mariners outfielder Ken Griffey Jr’s lifelong vendetta against New York Yankees; photo (M)

Table 4.8: Example outputs after removing training examples and retraining. Our method is able to correct some instances that the NER tagger (oracle) approach misses.

Original Output	Contrast
There is a high-priced coffee shop in the **City centre**. It is called Fitzbillies and it is family friendly, but it does have a 1 out of 5 rating.	There is a high-priced English coffee shop in the riverside area. It is called Fitzbillies and it is family friendly, but it does have a 1 out of 5 rating.
Browns Cambridge is coffee shop with low customer rating. It serves Chinese food. They are located in Riverside near the Crowne Plaza Hotel.	Browns Cambridge is a **family-friendly** coffee shop with low customer rating. It serves Chinese food. They are located in Riverside near the Crowne Plaza Hotel.

Table 4.9: Examples of contrasts used for the E2E setup. Semantic errors in the output are shown in **red**. The first example contains a hallucinated location (City center) that is not consistent with the location in the MR (riverside area). The second example shows a case where a slot that is present in the MR is omitted from the output (family-friendly).

4.2.2 Semantic Errors in the E2E dataset

4.2.2.1 Experimental Setup

In order to show that our error attribution approach works beyond text summarization, we also evaluate on the E2E dataset (Novikova et al. 2017), a popular benchmark for generating natural language descriptions from structured meaning representations (MRs). Prior work has shown that up to 40% of the E2E dataset contains some form of semantic noise, and models trained on this dataset tend to generate output that is not faithful to the input MR. They tend to either omit information in the MR or hallucinate new information that is not present in the MR (Dušek, Novikova, et al. 2020). In order to improve the semantic correctness of models trained on the E2E dataset, Dušek, Howcroft, et al. (2019) handcrafted rules to fix errors in the dataset based on manual analysis of hundreds of samples.

We study whether error attribution methods can be used to automatically identify noisy instances in the E2E training data, given just a few examples of generations with semantic errors. In particular, we select examples where the output contains a semantic error and then minimally edit the output to make it consistent with the MR, as shown in Table 4.9. We treat the manually cleaned dataset from Dušek, Howcroft, et al. (2019) as the oracle and measure how accurately error attribution methods are compared to this oracle. In particular, any training instances that were fixed by the manual rules from Dušek, Howcroft, et al. (2019) are treated as errors that the attribution methods should identify. We expect good attribution methods to reliably identify noisy training instances, which, when removed, can lead to models with improved semantic correctness without a drop in overall performance.

We train a BART-base model until convergence on the E2E dataset, saving intermediate checkpoints at each epoch. We use a learning rate of $1e-4$ and an effective batch size of 128. We then find 5 (random) descriptions from the validation set that contain semantic errors according to handcrafted rules from Dušek, Howcroft, et al. (2019) and use these examples as $\mathcal{D}_{\text{Err}}$ for error attribution. We post-edit the descriptions in $\mathcal{D}_{\text{Err}}$ to fix semantic errors for our contrast set, as shown in Table 4.9.[2] We update checkpoint 1 on $\mathcal{D}_{\text{Err}}$ for five gradient steps with a learning rate of $1e-5$. For distilling our contrastive influence estimates (S^c_{grad}), we take the top 500 scored training examples according to S^c_{grad} as positive examples and the bottom 500 scored examples as negative examples, and we finetune Electra (Clark et al. 2020) for 5 epochs with early stopping, with a learning rate of 2e-5 and a batch size of 8.

We expect a successful error tracing method to improve semantic correctness when we remove the error set $\mathcal{D}$. Therefore, we finetune a BART-base model after removing $\mathcal{D}$ identified by each method and compare the semantic error rate (SemErr) to the baseline. We use the scripts from Dušek, Howcroft, et al. (2019) to compute SemErr. For the oracle upper bound, we remove all training instances that would be corrected by the handcrafted rules from Dušek, Howcroft, et al. (2019) and re-train a BART-base model on the remaining training set.

4.2.2.2 Results

Table 4.10 shows the retrieval results for the different approaches on the E2E dataset. We treat the set of training instances for which the handcrafted rules from Dušek, Howcroft, et al. (2019) fire as the ground truth to measure auPR and auROC. Among the error tracing approaches, we find

[2]Note that unlike Dušek, Howcroft, et al. (2019) who use handcrafted rules to fix input MRs such that they match the description, we keep the MR unchanged and post-edit the description.

Method	AuPR	AuROC
Random	50.49	50.39
BM25	53.11	54.80
BartScore	52.87	54.24
TracIn	65.79	62.54
Ours	**71.60**	**65.34**

Table 4.10: Retrieval results on the E2E dataset. We see that our approach substantially improves upon prior work.

that BM25 and BartScore do not perform much better than the random baseline, while TracIn does substantially better. We see that our method does drastically better than all other methods in terms of auPR and auROC. Given these results, we perform the re-training experiments for only TracIn and our approach, along with the baseline and oracle.

Method	SemErr	ROUGE-L	BLEU
Baseline	6.08	53.42	33.81
Oracle	1.43	54.44	35.42
TracIn	5.08	54.10	34.90
Ours	2.76	54.19	35.19

Table 4.11: Semantic Error Rate (SemErr) for retrained models after removing erroneous examples identified by each method. We see that our approach does considerably better than TracIn.

Table 4.11 shows the results of retraining after removing erroneous training instances identified by each method. For a fairer comparison, we remove the same number of instances as identified by the oracle. We see that our method reduces the relative SemErr of the baseline by almost 55% compared to a more modest 16% reduction for TracIn. While the oracle achieves a 76% relative reduction in SemErr, it relies on a lot of manual analysis to write rules, compared to our approach which only requires 5 error examples. Furthermore, we see that the ROUGE-L and BLEU scores for our approach are comparable to the oracle system.

4.3 Related Work

Recent work in abstractive summarization has shown that state-of-the-art systems suffer from generating information that is inconsistent with the source article, despite their improved success in producing fluent summaries (Falke et al. 2019; Durmus, H. He, et al. 2020; Lux et al. 2020; Wilber et al. 2021). Improving faithfulness of summarization systems is essential for deploying these systems in real-world scenarios, as such recent work has studied methods to improve the faithfulness of abstractive summarization systems (Matsumaru et al. 2020; Z. Zhao et al. 2020; Y. Dong, S. Wang, et al. 2020; Goyal and Greg Durrett 2021; Xu et al. 2020; S. Chen et al. 2021; Zhu et al. 2021). For example, Goyal and Greg Durrett (2021) train summarization systems by modifying the training objective to maximize the likelihood of the subset of summary tokens that are considered faithful according to their factuality detection model. Z. Zhao et al. (2020) specifically target hallucination of quantities in generated summaries and train a verification model that they use to re-rank summaries such that summaries containing quantities consistent with the source article are up-ranked. Kang and T. B. Hashimoto (2020) propose a method to adaptively remove high log loss examples and show that this method achieved better faithfulness scores compared to the baseline. Although these methods have shown improvements over the compared baselines, unlike our work, they do not take extractiveness of the generated outputs into account, and as we showed in §2.2. In contrast, the faithfulness mitigation strategy that we propose in §4.1 shows that by taking into account extractiveness of generated summaries, we can build a dynamic system that can fall back to more extractive summaries when unsure leading to an overall mode faithful summarizer that still improves the *effective faithfulness* measure that we proposed in §2.2.

Since prior work has shown that some of the faithfulness errors observed in generation systems can be due to the noise in the dataset (Maynez et al. 2020; Kang and T. B. Hashimoto 2020), we also explore how the error attribution method we proposed in §3.1 can be used to mitigate errors in text generation systems. Influence function based methods Koh and Liang 2017 are closest to our approach, as they seek to understand how removing data impacts model predictions, often in classification settings X. Han et al. 2020. While there have been substantial improvements upon the original Taylor approximation-based method (Koh and Liang 2017) via the use of multiple checkpoints (Pruthi et al. 2020) and modifications to the hessian approximation (Hara et al. 2019; Schioppa et al. 2022), they can be brittle and recent work has shown that they can underperform lexical similarity baselines (Akyürek et al. 2022). Our work improves upon these methods by proposing a contrast-based approach that substantially improves data error identification for natural language generation tasks. For error tracing, there are embedding and similarity-based methods that seek to find examples that are similar to a given test example or error (Rajani et al. 2020; W. Yuan et al. 2021). However, we find that although these methods often improve upon influence-based estimates and are useful for interpreting errors, they still do not achieve high enough precision and recall to substantially improve downstream properties such as hallucination rates. We show in §4.2 that our approach outperforms all prior work in reducing faithfulness errors in generation systems. We note that our approach of using error attribution for data cleaning can complement other solutions to this problem, such as data augmentation via self-training (Kedzie and Mckeown 2019), or the growing literature on modeling-based solutions to this problem, including using information extraction (Z. Cao et al. 2018) or a QA model (Nan, Nogueira dos Santos, et al. 2021).

4.4 Conclusion and Limitations

In §4.1 we explore controlling abstractiveness as a way of improving overall faithfulness. In particular, we generate several summaries at different levels of abstractiveness and train a selector model that can select the most abstractive and faithful summary from the set of candidate summaries. We show that our approach can improve faithfulness over the baseline system while also being more abstractive on average. Furthermore, we show that, unlike prior work that aims at improving the faithfulness of summarization systems, our work does not simply rely on increased extractiveness to improve faithfulness. We show that our work is able to consistently improve *effective faithfulness* across datasets. Furthermore, we show that by adjusting the weight between the precision and recall of our selector, we can get systems operating at different levels of extractiveness while still consistently improving *effective faithfulness*. Our work shows that building summarization systems that are more aware of the faithfulness-abstractiveness trade-off can be a viable way to mitigate faithfulness errors, and we encourage future work to explore this further.

In §4.2 we explore whether error attribution can be used to produce cleaner datasets that lead to fewer faithfulness errors in model generation. In particular, we compare our proposed error attribution method against prior work on two real-world generation tasks – summarization on the NYT dataset and MR to text generation on the E2E dataset. We show that our method outperforms prior error attribution approaches by a significant margin on both tasks and leads to substantial reductions in faithfulness errors. On the NYT summarization task, our error attribution method leads to a comparable reduction in entity hallucinations as a strong oracle system using a state-of-the-art NER tagger (spaCy), despite only using a handful of error examples. Similarly, on the E2E dataset, despite only using just a few error examples, the SemErr of our approach is not too far behind the oracle system, which required a significant amount of manual analysis and handwritten rules. Overall, our work shows that our error attribution method can be a useful tool to reduce dataset noise and build better generation systems.

4.4.1 Limitations

Our adaptive selector approach to improving faithfulness relies on the ability to generate summaries with varying levels of extractiveness. In our work, we actually used summarizers that are trained on subsets of the dataset at different levels of extractiveness, which means that we had to train several summarizers along with the selector model. Ideally, we would have a single summarization system with controllable generation for the level of extractiveness. While we did try to train a controllable generation model using the approach proposed by Keskar et al. (2019), we were unable to get that model to reliably generate summaries of different levels of extractiveness. However, we note that as the literature on controllable generation improves, we should be able to make our approach more efficient by replacing the existing summarizers with a single model.

Our approach to improving faithfulness by cleaning datasets using error attribution methods is based on the premise that faithfulness errors observed in generation systems are due to noise in the dataset. While there is substantial evidence for this from prior work, and our methods achieve remarkable results in the datasets we used, it's possible that the utility of our approach could drop in cases where we have clean, curated datasets. It's possible that certain faithfulness errors made by the model could be due to spurious patterns learned by the model that do not generalize well. In such cases, it's unclear whether using our error attribution approach to remove training instances would alleviate the problem. However, as most large-scale datasets in natural language generation tend to be sourced from the internet, it's inevitable that these datasets will likely contain at least a few erroneous examples that could lead to undesirable model generations. Therefore, we believe that our approach to using error attribution to clean datasets is still a valuable method to improve the faithfulness of model generations.

Chapter 5: Benchmarking Large Language Models for Summarization

Our work thus far has focused on finetuned LMs since supervised finetuning of pretrained LMs using large-scale summarization datasets has been the recipe for success for the current state-of-the-art in text summarization (Yang Liu and Lapata 2019; Lewis et al. 2020; J. Zhang et al. 2020; Yixin Liu, P. Liu, et al. 2022). Since language model pretraining has been such a huge component for success in all areas of natural language processing, recent work has spent considerable effort in scaling up language models both in terms of parameters and dataset size (Brown et al. 2020; Chowdhery et al. 2022; Bommasani, Hudson, et al. 2021). Scaling up parameters and data to produce this new generation of Large Language Models (LLMs) leads to a key distinctive quality over the previous generation of pretrained language models. With some additional instruction tuning, these models do not require task-specific finetuning and can be prompted in a zero-shot or few-shot manner to solve a task (Brown et al. 2020; Chowdhery et al. 2022; Yushi Bai et al. 2022).

In the zero-shot setting, prompting presents the LLM with inputs (e.g., news articles) and a natural language instruction (e.g., “summarize this news article in three sentences”) and solicits an output by having the LLM generate a completion, representing the answer. When few-shot training examples are provided, LLMs have the ability to learn "in context" from these examples (Brown et al. 2020). In-context learning prepends training input-output pairs along with the same style of instruction to the testing input. While initial LLMs produced dramatically good results using zero-shot and few-shot prompting, their performance was still lagging behind the state-of-the-art for many natural language processing tasks (Brown et al. 2020).

Recently, instruction-tuning has emerged as an effective way to improve LLM prompting performance (Sanh et al. 2021; Y. Wang et al. 2022; Ouyang et al. 2022). In this approach, a diverse set of natural language processing tasks are reformulated into the prompting format, and the LLM’s parameters are updated for these tasks through supervised finetuning or reinforcement learning.

This multi-task tuning of the model's parameters allows it to generalize better to new instructions at test time and leads to stronger zero-shot and few-shot performance across a variety of standard natural language processing tasks.

In this chapter, we aim to benchmark LLMs for their ability to perform text summarization in zero-shot and few-shot settings. In § 5.1. Recent work (Goyal, J. J. Li, and Greg Durrett 2022) shows that the instruct-tuned GPT-3 Davinci model (Ouyang et al. 2022) is better than finetuned LMs for new summarization but does not show the design decisions that contribute to the improved performance. In our work, we carry out a more comprehensive benchmark on ten different LLMs to understand the effect of model scale, in-context learning, and instruction tuning. In particular, we carry out human evaluations, since automated metrics are unreliable (see §2.1), across two popular news summarization benchmarks in order to assess what design decision of LLMs contributes the most to their strong performance in news summarization. In addition, we also benchmark LLMs against human writers to understand how far we are from human-level performance in news summarization. Overall, we find that summaries generated by the best instruction-tuned LLM perform as well as human-written summaries in faithfulness, coherence, and relevance, though they differ stylistically.

In § 5.2, we explore whether the strong performance we observed in news summarization translates to other summarization tasks. In particular, we explore whether state-of-the-art LLMs can generate coherent and faithful summaries for narratives. Prior work has explored several different types of narratives such as novel summarization (Mihalcea and Ceylan 2007; Ladhak, B. Li, et al. 2020; Kryściński, Rajani, et al. 2021), movie screenplay summarization (Papalampidi et al. 2020; C. Zhao et al. 2022), and TV series screenplay summarization (M. Chen et al. 2022; C. Zhao et al. 2022). Finetuning LMs for these tasks has seen limited success due in part to the difficulty of summarizing long documents, and current state-of-the-art systems for these tasks tend to generate summaries of poor quality (Goyal, J. J. Li, and Greg Durrett 2022). In order to disentangle the effects of difficulties in narrative understanding vs. long document comprehension, in our work, we focus on the summarization of short stories. We first compile a new benchmark dataset of short stories collected from Reddit. We then benchmark three state-of-the-art LLMs for zero-shot summarization on this dataset, asking experts in the field to rate these systems for faithfulness. We find that unlike news summarization, where LLMs were almost perfect in terms of faithfulness, LLMs are unable to generate faithful summaries for this task. We then show that getting reliable evaluations via crowd-sourcing for this task is difficult, as annotators get very low agreements with experts. We then present a model-in-the-loop approach to faithfulness evaluations, showing that it leads to more reliable crowd-sourced annotations.

5.1 Benchmarking LLMs for News Summarization

This is based on T. Zhang, Ladhak, et al. (2023), which was work done in collaboration with researchers from Stanford University. I was an equal contribution first author; ordering was decided via coin flip.

Large language models (LLMs) have shown promising results in zero-/few-shot tasks across a wide range of domains (Chowdhery et al. 2022; Yushi Bai et al. 2022; Brown et al. 2020; Susan Zhang et al. 2022) and raised significant interest in their potential for automatic summarization (Goyal, J. J. Li, and Greg Durrett 2022; Yixin Liu, Alexander R. Fabbri, et al. 2022). However, the design decisions contributing to its success in summarization remain poorly understood. While prior work has shown that LLMs outperform the previous state-of-the-art, it remains unclear whether their outputs are comparable to human writers. Examining these questions is crucial for advancing future research in automatic summarization.

To answer the first question, we perform a systematic human evaluation of ten diverse LLMs on the news summarization task. Our evaluation identifies instruction tuning to be the key to zero-shot summarization capability. In contrast, self-supervised learning alone cannot induce strong summarization performance in the zero-shot setting (Figure 5.1). In fact, we find that even a 350M parameter instruction-tuned GPT-3 can perform on par with the 175B parameter GPT-3.

To benchmark LLMs, we rely on the standard CNN/DM (Hermann et al. 2015) and XSUM benchamrks (Narayan et al. 2018) but find that the poor quality of the existing reference summaries causes several issues for our evaluation. The reference summaries in these benchmarks are of such

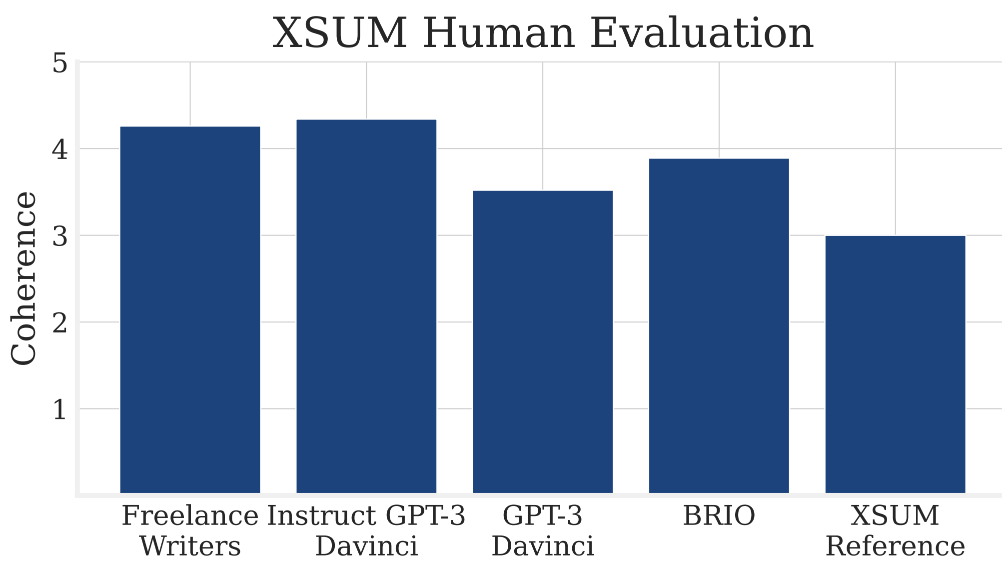


Figure 5.1: Annotator ratings of summary coherence on a 1 to 5 Likert scale for a selected set of models. We see that the instruct-tuned variant of GPT-3 performs on par with human writers.

a poor quality that human annotators judge them to be worse than the outputs of most automatic summarization systems (Figure 5.1). Furthermore, when we use these references to evaluate systems using automated evaluation metrics, the poor quality of the references leads to a low correlation between metric scores and human judgment. Not only does this make evaluation difficult, but it also means that systems that use these references for supervision, whether through finetuning or few-shot prompting, will experience a degradation in performance, making a fair comparison across these approaches more challenging. Furthermore, these poor-quality references do not represent how humans would write summaries and, therefore, do not help us understand how close LLMs are to producing human-quality summaries.

To address the quality issues of reference summaries and to better understand how LLMs compare to human summary writers, we recruit freelance writers from Upwork[1] to write summaries 100 random articles from the test set of CNN/DM and XSUM. Comparing the best performing LLM, Instruct Davinci, to the freelance writers, we find that the Instruct Davinci summaries are much more extractive. By manually annotating the summarization operations (Jing and K. McKeown 2000) used in these summaries, we find that Instruct Davinci paraphrases much less frequently than

[1] `https://www.upwork.com`

the freelance writers, although it is able to combine copied segments from different sentences in a coherent manner.

Given the stylistic differences between LLMs and human writers, we recruit annotators to compare the Instruct Davinci summaries to those written by freelance writers. On aggregate, we find that Instruct Davinci is rated as comparable to the freelance writers. However, analysis of individual annotators reveals that each annotator has a varying and stable preference for either Instruct Davinci or freelance writers.

Overall, our work makes the following key contributions. First, we identify instruction tuning and Reinforcement Learning from Human Feedback (RLHF), rather than model scale, as the key to LLMs' summarization capability. Second, we show that humans judge the reference summaries used in XSUM to be worse than model-generated summaries. Third, to address the issue of low-quality references, we collect higher-quality summaries from freelance writers, and we show that the best LLM is rated as comparable to freelance writers, according to human annotators. In combination, these results call into question recent claims made about LLM summarization. In particular, summarization progress cannot be measured using reference-based metrics applied on XSUM. Furthermore, the question of whether finetuned, few-shot, or zero-shot models perform better remains an open question due to the poor quality of references in the training data. To encourage future work on improved evaluation, we release our high-quality reference summaries written by freelance writers, along with the evaluation data across 18 models and two datasets.[2]

5.1.1 Human Evaluation on News Summarization Benchmarks

In this section, we use human evaluation to systematically benchmark a diverse set of 10 LLMs on the news summarization task. We observe that instruction tuning is the key to the strong summarization capability of LLMs, and the low-quality reference summaries in current benchmarks may underestimate few-shot or finetuning performance.

[2] `https://github.com/Tiiiger/summ_dataset`.

5.1.1.1 Experimental Setup

Model	Creator	Size	Instruction Tuning	Reference
GPT-3 davinci v1 GPT-3 curie v1 GPT-3 ada v1	OpenAI	175B 6.7B 350M	✗	Brown et al. (2020)
InstructGPT davinci v2 InstructGPT curie v1 InstructGPT ada v1	OpenAI	175B 6.7B 350M	✓	Ouyang et al. (2022)
OPT 175B	Meta	175B	✗	Susan Zhang et al. (2022)
GLM	Tsinghua University	130B	✗	Du et al. (2021)
Cohere xlarge v20220609	Cohere	52.4B	✗	Cohere (2022)
Anthropic-LM v4-s3	Anthropic	52B	✓	Yushi Bai et al. (2022)

Table 5.1: The list of LLMs we benchmarked on CNN/DM and XSUM with human evaluation.

Data. We conduct our human evaluation on CNN/DM and XSUM by sampling 100 examples from each validation set, respectively. For the few-shot in-context learning settings, we sample five examples from the training set as demonstration examples. Due to the limited context window, we sample five articles between 50 and 150 tokens in length according to the GPT-2 tokenizer. For XSUM, we find that uniform sampling occasionally results in unreadable articles due to data preprocessing, so we manually pick five samples from the training set.

Model Details. We consider 10 LLMs across different pretraining strategies and model scales.[3] Table 5.1 lists the details of the LLMs we consider for our evaluation. Due to limited computational resources and model access, we benchmark all models in the five-shot setting but only benchmark the three OpenAI GPT-3 models and the three OpenAI instruction-tuned GPT-3 models in the zero-shot setting. For CNN/DM, we solicit LLM summaries with the following prompt template:

[3]We note that the training details of instruction-tuned GPT-3 models may differ from those mentioned in the publication and are inferred by us based on the API naming scheme.

```
Article: [Article text]
Summarize the article in three sentences.
Summary:
```

For XSUM, we modify the prompt template to ask the model to summarize in a single sentence to match the style of the reference summaries. We sample all LLMs with temperature 0.3 following prior work J. Wu et al. 2021. To contextualize our LLM benchmarking results, we also evaluate two state-of-the-art finetuned LMs: Pegasus (J. Zhang et al. 2020) and BRIO (Yixin Liu, P. Liu, et al. 2022).[4] We decode the finetuned LMs using a beam size of 5 following prior work (Lewis et al. 2020). In addition, we also evaluate the existing reference summaries in the CNN/DM and XSUM validation sets.

Human Evaluation Protocol. We recruit annotators from Amazon Mechanical Turk, compensating them at California minimum wage of $15.00/hr using conservative time estimates as recommended by Whiting et al. (2019). Three annotators evaluated each model summary, and we took the average across the three annotators as the overall score for each summary. Our annotators evaluate each summary based on three criteria: faithfulness, coherence, and relevance. We define these terms and collect data according to the guidelines in Alexander R Fabbri et al. (2020). Coherence and relevance ratings are collected on a 1 to 5 Likert scale, while faithfulness ratings are collected as binary ratings due to its binary nature. Unlike Alexander R Fabbri et al. (2020), we omit to evaluate fluency because we find LLM outputs highly fluent. The complete annotation guidelines are included in our code release.

5.1.1.2 Evaluation Results

Table 5.2 shows the human evaluation results for both benchmark datasets.[5] We now discuss the two main takeaways that we observe in our evaluation results.

[4] We use the trained checkpoints made available in the Hugging Face model hub.

[5] We note that the 350M GPT-3 model consistently generates empty outputs for XSUM, so we omit it from the human evaluation.

Setting	Models	CNN/Daily Mail Faithfulness	CNN/Daily Mail Coherence	CNN/Daily Mail Relevance	XSUM Faithfulness	XSUM Coherence	XSUM Relevance
Zero-shot language models	GPT-3 (350M)	0.29	1.92	1.84	0.26	2.03	1.90
	GPT-3 (6.7B)	0.29	1.77	1.93	0.77	3.16	3.39
	GPT-3 (175B)	0.76	2.65	3.50	0.80	2, .78	3.52
	Ada Instruct v1 (350M*)	0.88	4.02	4.26	0.81	3.90	3.87
	Curie Instruct v1 (6.7B*)	0.97	**4.24**	**4.59**	**0.96**	4.27	**4.34**
	Davinci Instruct v2 (175B*)	**0.99**	4.15	**4.60**	**0.97**	4.41	**4.28**
Five-shot language models	Anthropic-LM (52B)	0.94	3.88	4.33	0.70	**4.77**	4.14
	Cohere XL (52.4B)	**0.99**	3.42	4.48	0.63	**4.79**	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	OPT (175B)	0.96	3.64	4.33	0.67	**4.80**	4.01
	GPT-3 (350M)	0.86	3.73	3.85	-	-	-
	GPT-3 (6.7B)	0.97	3.87	4.17	0.75	4.19	3.36
	GPT-3 (175B)	**0.99**	3.95	4.34	0.69	4.69	4.03
	Ada Instruct v1 (350M*)	0.84	3.84	4.07	0.63	3.54	3.07
	Curie Instruct v1 (6.7B*)	0.96	**4.30**	4.43	0.85	4.28	3.80
	Davinci Instruct v2 (175B*)	**0.98**	4.13	4.49	0.77	**4.83**	**4.33**
Finetuned language models	Brio	0.94	3.94	4.40	0.58	4.68	3.89
	Pegasus	0.97	3.93	4.38	0.57	4.73	3.85
Existing references	-	0.84	3.20	3.94	0.37	4.13	3.00

Table 5.2: Human evaluation results for zero-shot and five-shot LLMs, finetuned LMs, and reference summaries. We bold all entries that are statistically similar to the best numbers in each column.

Instruction-tuned models have strong summarization ability. Across the two datasets and the three evaluation aspects, we find that the zero-shot instruction-tuned GPT-3 models perform the best overall, especially Instruct Curie and Davinci. Compared to the finetuned LMs (e.g. Pegasus), Instruct Davinci achieves higher coherence and relevance scores (4.15 vs. 3.93 and 4.60 vs. 4.40) on CNN and higher faithfulness and relevance scores (0.97 vs. 0.57 and 4.28 vs. 3.85) on XSUM, which is consistent with recent work (Goyal, J. J. Li, and Greg Durrett 2022). In contrast to instruction tuning, we find model scale to be less important for summarization performance. Even the largest 175B model often ignores the instruction and generates irrelevant content while being outperformed on coherence and relevance by the much smaller Instruct Ada model.

In the five-shot setting, the non-instruction-tuned LLMs can improve their summarization performance through in-context learning. For example, several non-instruction-tuned LLMs perform on par with their instruction-tuned counterparts on faithfulness scores in CNN/DM and coherence scores on XSUM. However, we still find the instruction-tuned LLMs perform better for other evaluation aspects.

Article: https://www.dailymail.co.uk/femail/article-3052124

Reference summary:

Rare tote is so exclusive that it can't even be brought in store. Experts say it's an extremely good investment as only increases in value. Kim Kardashian and Victoria Beckham are both fans of Hermes.

0-shot GPT-3 Davinci summary:

The most expensive handbag in the world is a pink Hermes Birkin bag, which has been put on sale for £140,000. The bag is so exclusive that it can't even be brought in store and experts say it's an extremely good investment. With the same amount of money that the handbag costs, you could splash out on an Aston Martin DB9 Volante.

What is the main idea of the article?

0-shot Instruct Davinci summary:

The article discusses a pink Hermes Birkin bag that is on sale for £140,000. The bag is made-to-measure and embellished with rubies. The article also discusses how the bag is a collector's item and an investment.

Freelance writer summary:

Hermes is a brand rich in history and value, so much so that experts are saying their new £140,000 handbag is an extremely good investment. The bag, studded in rubies, is only the latest in Hermes Birkin's long line of highly expensive handbags for highly stylish women, which include celebrities and royalty.

Figure 5.2: Summaries generated by GPT-3 models (Section 5.1.1) compared to a summary written by a freelance writer (Section 5.1.2) for an article from the CNN/DM dataset. We find that the instruction-tuned GPT-3 model can generate a higher-quality summary than the non-instruction-tuned variant. The reference summary from the CNN/DM dataset is not coherent, whereas the freelance writer's summary is both coherent and relevant.

Reference summaries in current benchmarks are extremely low quality. Our evaluation results show that reference summaries in the CNN/DM and XSUM datasets are of poor quality. First, we observe that most automatic summarization systems score better than the reference summaries across all three evaluation aspects. Second, using these existing reference summaries as in-context examples for the few-shot setting actually degrades the performance of instruction-tuned models, leading them to generate summaries that are scored worse by human raters. For instance, on the XSUM dataset, incorporating five reference summaries into the context leads to a large drop in the faithfulness score of Instruct Davinci (from 0.97 to 0.77).

As a consequence, the low quality of existing reference summaries makes it difficult to reliably compare LLMs to both finetuned models and humans. The poor performance of finetuned models can be attributed to the low-quality references in the training data, and we may be underestimating the finetuning performance. Similarly, these low-quality references do not represent human performance on news summarization because they are created through heuristics. As a result, it's likely that the differences between instruction-tuned LLMs and human performance are likely overstated in Table 5.2.

Qualitative Examples. Figure 5.2 compares summaries of zero-shot GPT-3 Davinci, instruction-tuned GPT-3 Davinci, and the CNN/DM reference for an article from the CNN/DM validation set. We note that the zero-shot GPT-3 model is unable to follow the instructions and generate a good summary. After generating an initial summary, the model generates an additional question that is completely irrelevant to the instruction. This tends to happen quite often for non-instruction-tuned models. In addition to a failure to follow the instructions, the generated summary also contains a factual error, stating that the handbag mentioned is the most expensive in the world, which contradicts the original article. In contrast, the instruction-tuned GPT-3 model generates a summary that is both faithful and coherent. We also observe that the reference summary in the CNN/DM dataset is not coherent. The brand "Hermes" is not introduced until the end and its connection to the rest of the story is unclear. This is unsurprising as reference summaries in the CNN/DM dataset were originally bullet points accompanying the articles as opposed to a coherent summary paragraph.

5.1.1.3 Understanding Automated Evaluation Metrics

We compute system-level correlations against human ratings for eight popular automated evaluation metrics. For reference-based metrics we consider: Rouge-L (Lin 2004), METEOR (Banerjee and Lavie 2005), BertScore (T. Zhang, Kishore, et al. 2019), BLEURT (Sellam et al. 2020), and BARTScore (Hara et al. 2019). For reference-free metrics we consider: SummaC (Laban et al. 2022), QAFactEval (A. Fabbri et al. 2022), and BLANC (Vasilyev et al. 2020).

For CNN/DM, we observe that the reference-based automatic metrics have a moderate correlation with some aspects of human judgments, e.g., Rouge-L has a 0.72 Kendall's tau correlation coefficient with relevance in Table 5.3. Such a level of correlation is comparable to that reported

Metric	CNN/DailyMail Faithfulness	CNN/DailyMail Coherence	CNN/DailyMail Relevance	XSUM Faithfulness	XSUM Coherence	XSUM Relevance
Rouge-L	0.54	0.48	0.72	-0.27	0.71	0.30
METEOR	0.58	0.37	0.66	-0.22	0.68	0.38
BertScore	0.54	0.47	0.70	-0.23	0.70	0.30
BARTScore	0.56	0.34	0.65	-0.22	0.70	0.35
BLEURT	0.56	0.62	0.81	-0.08	0.67	0.41
SummaC	0.54	0.11	0.26	0.26	-0.41	-0.29
QAFactEval	0.64	0.16	0.35	0.55	0.16	0.37
BLANC	0.54	0.31	0.50	0.50	0.10	0.32

Table 5.3: System-level kendall's tau correlation of automated metrics with human scores across the different evaluation aspects.

in Alexander R Fabbri et al. (2020), which measures the correlation of automated metrics on evaluating finetuned LMs and even earlier neural summarization systems. Therefore, we conclude that on CNN/DM, automated reference-based evaluation metrics can still provide useful signals for relevance.

Table 5.3 shows Kendall's tau rank correlations between automated metrics and human judgments. We observe significantly different trends on CNN/DM and XSUM, so we will discuss them separately in the following paragraphs.

Studying the result more closely, we find that Rouge-L and human evaluation are more correlated when comparing within each model group. We plot Rouge-L over the relevance rating in Figure 5.3 as an example. First, we observe that Rouge-L still prefers finetuned LMs (green points on top of the plots) to LLMs, consistent with prior work (Goyal, J. J. Li, and Greg Durrett 2022). Despite this error, when only comparing LLMs with each other, we find that a larger than 0.05 Rouge-L difference usually translates to improved human evaluation.

On XSUM, reference-based evaluation metrics have a very low correlation with faithfulness and relevance since the reference summaries themselves are terrible for these aspects (Table 5.3; also see Maynez et al. 2020). With such low-quality references, we do not expect reference-based evaluation metrics to provide useful information about differences between the summarization systems.

In general, across both datasets, we find that reference-based evaluation metrics correlate better

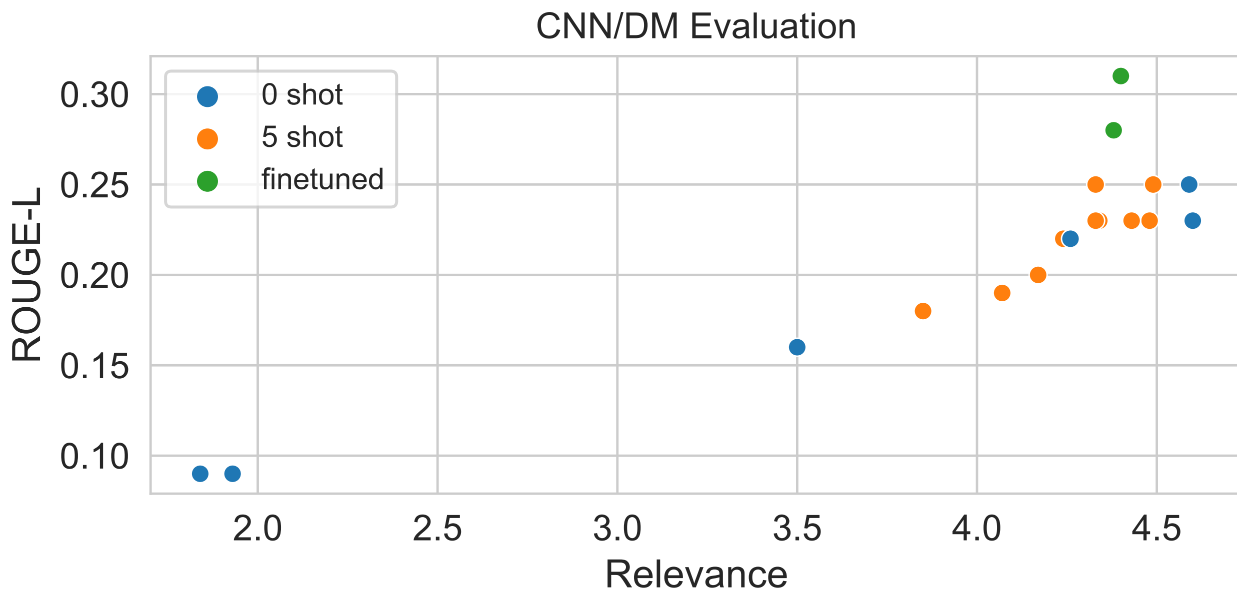


Figure 5.3: System-level Rouge-L vs. annotator rated relevance scores.

with human judgments on the aspects for which reference summaries also have better scores (e.g. CNN/DM relevance, XSUM coherence). This points to the important role of the quality reference summaries for reference-based evaluation metrics, as previously observed in machine translation (Freitag et al. 2020). Reference-free evaluation metrics are less handicapped by the low quality of existing references, but they are mostly geared toward measuring faithfulness. Even BLANC, which is designed to measure overall summary quality, correlates best with faithfulness and is much worse for relevance and coherence.

5.1.2 Comparing Summarization Systems to Freelance Writers.

In Section 5.1.1, we saw that the low-quality reference summaries in current news summarization datasets make studying and benchmarking LLMs difficult. In this section, we address this by recruiting freelance writers from Upwork to collect better-quality summaries for news articles. With these newly collected references, we aim to answer two important questions. First, we would like to know whether the best LLM has reached human-level performance for news summarization and how the summaries written by the LLM differ from the ones written by humans. Second, we want to understand how well reference-based evaluation metrics correlate with human judgments once we compute them using higher-quality reference summaries.

5.1.2.1 Experimental Setup

In this section, we describe the recruitment process and instructions for the summary writing task.

Data. For the data used in our study, we select 50 articles from each of the CNN/DM and XSUM evaluation sets described in Section 5.1.1.1 and assign each article to three writers. During the XSUM dataset construction, the authors removed the first sentence of the article and used that as the reference summary for the rest of the article. The articles without the first sentence are usually very difficult to follow as key information that is crucial for understanding the article has been stripped. Therefore, when collecting the new references, we take the original, unaltered article with the first sentence intact. In order to have a fair comparison, we also prompt the LLM with the original, unaltered article.

Writer recruitment. We recruit six writers with previous experience in writing blog posts, landing page introductions, or product descriptions from the freelance work platform Upwork. After conducting a qualification round by asking writers to summarize five articles, we selected the best writers according to the faithfulness, coherence, and relevance of their summaries. Through an initial pilot study, we estimate that the time required to summarize a CNN/DM or XSUM article is around 12 to 15 minutes. Therefore, we pay our writers \$4 for every article they summarize to following the recommended practice (Whiting et al. 2019). We based the assignments on writers' availability, with the most prolific writer summarizing 100 articles and the least prolific writer summarizing 35 articles.

Summary writing instructions. For the summary writing instruction, we instruct our writers to summarize each article in around 50 words.[6] To provide some grounding for the summary writing

[6]We conducted an initial study to pilot instructions and found that instructing writers with a sentence limit often resulted in summaries that differ significantly in length.

Model	Faithfulness	Coherence	Relevance
Freelance Writer	0.93	4.39	4.26
Zero-shot Instruct Davinci	0.98	4.26	4.40
Reference Summaries	0.64	3.59	3.45

Table 5.4: Amazon Mechanical Turk evaluation results of the freelance writer summaries. Results of zero-shot Instruct Davinci and reference summaries are taken from Table 5.2 after averaging the corresponding ratings.

task, we ask the writers to summarize as if they are writing a newsletter to update their readers on the news.

LLM Summaries Generation. Recently, Yixin Liu, Alexander R. Fabbri, et al. (2022) showed that length is a confounding factor in summarization human evaluation. To control for this potential length confound, we modify the zero-shot prompt in Section 5.1.1.1 to elicit summaries that are around 50 words, which is the same word limit provided to the freelance writers. Through initial pilots, we found that the Instruct Davinci model does a poor job of following length constraints and consistently produces summaries that exceed the given word limit. Therefore, we intentionally prompt the Instruct Davinci model with a 25 word limit to produce summaries with an average length of 50 words. With this new prompt, we generate the summaries using the same hyperparameters described in Section 5.1.1.1.

Quality Control. To verify the quality of the summaries written by freelance writers, we evaluate a random subset of 100 summaries using the same annotation scheme in Section 5.1.1.1 on Amazon Mechanical Turk. Table 5.4 shows the results of this evaluation. We see that the freelance writer summaries are of a much higher quality than the original reference summaries in CNN/DM and XSUM. In addition, we see that the difference between the freelance writers and Instruct Davinci in this evaluation is small. Next, we perform more targeted evaluations to compare summaries written by freelance writers and Instruct Davinci.

5.1.2.2 Paired Comparison between LLM and Freelance Writers.

Using this higher-quality set of human-written summaries, we benchmark the best LLM (Instruct Davinci) by having annotators directly compare the model-generated summaries against the human-written summaries.

Comparing Stylistic Differences. Despite the similar performance in our quality control study, we find that LLM summaries and freelance writer summaries have distinctive styles. Figure 5.2 shows an example summary written by the freelance writer. Compared to the LLM-generated summary, we find the freelance writer summary often contains more paraphrasing and copies less from the article.

To illustrate this stylistic difference, we measure two extractiveness measures, coverage and density, following Grusky et al. (2018). Coverage is defined as the percentage of words in the summary that are also present in the article; density is defined as the average length of the continuous text spans in the summary that are copied from the article. Our analysis shows that the coverage and density for summaries generated by Instruct Davinci are 0.92 and 12.1, whereas those for the writers' written summaries are 0.81 and 2.07. These measures show that the summaries generated by Instruct Davinci are highly extractive, whereas the summaries written by freelance writers are much more abstractive.

To have a fine-grained understanding of these stylistic differences, we manually analyze the distribution of "cut and paste operations" in these two sets of summaries. Jing and K. McKeown (2000) identify a set of "cut and paste" operations for reusing text from the article, including sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing, and generalization or specification. On top of these operations, we additionally include a sentence copy operation to account for summary sentences that are directly copied from the article. Using this guideline, we manually annotate 10 randomly sampled summary pairs written by Instruct Davinci and the freelance writers.

Figure 5.4 reports the distribution of the cut-and-paste operations, showing the fraction of sentences that contain each operation. First, we observe that the freelance writer summaries use lexical paraphrasing and generalization/specification much more frequently than the Instruct Davinci generated summaries. Because both operations often involve using novel words that are not present in the article, this matches with the fact that the freelance writer summaries have lower coverage (0.81 vs. 0.92) than the Instruct Davinci summaries. Second, we find that sentence combination is a common strategy used by both the freelance writers and Instruct Davinci. Third, we find that freelance writers never copy an entire sentence directly from the article, but Instruct Davinci does this more frequently. Similarly, the Instruct Davinci model does not attempt to produce generalizations/specifications, whereas human writers do this more frequently.

In conclusion, we find that Instruct Davinci summarizes in a very different style than human writers. We emphasize here that the freelance writers write in an abstractive style despite the fact that we have not explicitly instructed them to do so. We also observe similarly abstractive styles across the six freelance writers.

Comparing Human Preference. We now return to our original goal of understanding whether the quality of LLM-generated summaries is on par with the human-written ones. Since we saw in the previous section that human-written and LLM-generated summaries have very distinctive styles, we would now like to understand whether or not human readers have a preference for a particular style of summary. In the following paragraphs, we describe our annotation design and recruitment process.

We conduct a blinded pairwise comparison between the best LLM (Instruct Davinci) and freelance writers, similar to the evaluation in Goyal and Greg Durrett (2020). The annotators have the option of selecting either of the two summaries if they judge one to be better than the other or selecting that both summaries are equally good. We release the full annotation instructions along with the code release for this project.

In order to compare the best LLM with the freelance writers, we focus on two aspects. First, we solicit annotators' overall preference, which balances the multiple quality aspects such as faithfulness, coherence, and relevance. Second, we solicit a more targeted measure of informativeness by asking the annotators to compare the number of facts in each summary. For the informativeness measure, we are motivated by the hypothesis that a more abstractive writing style could allow writers to pack more information into the summary with the same length constraint. While it is also interesting to compare summary coherence and relevance, we omit them because annotators were unable to differentiate these aspects from the overall preference in a pilot study.

For our recruitment process, we recruit five additional annotators through Upwork and retain one writer who participated in the previous round of summary writing.[7] We carry out a qualification round and reject annotators whose ratings differ significantly from the authors' on a set of control questions for informativeness. We give each annotator the same set of 100 summary pairs, where the average length of the freelance writer summaries and the Instruct Davinci summaries are 53.2 and 52.0respectively.

Figure 5.5 shows the results of the paired comparison. While we hypothesized that the more abstractive writing style could lead to more informative summaries, we did not find a significant effect in our annotator pool, who rate the more abstractive summaries to be more informative only 51.1% of the time. On the informative question, our annotators reached a moderate agreement (Krippendorff's alpha is 0.32), validating our annotation instruction and recruitment process. Moving onto the more subjective overall preference, we find that our annotators equally prefer the freelance writer summaries and the Instruct Davinci summaries. However, a closer analysis shows that there is significant variability in individual annotators' preferences, and the inter-annotator agreement is low (Krippendorff's alpha is 0.07). This suggests that in terms of quality, the LLM-generated summaries are similar to those written by freelance writers and that the differences we observe between annotators come down to individual stylistic preferences.

One example of such a stylistic preference is seen in the results from annotator 1, who also participated in the first round of summary writing. Like other writers, annotator 1 summarizes in an abstractive style (2.5 density and 0.86 coverage). However, annotator 1 prefers Instruct Davinci

[7]Other annotators unfortunately had to drop out after the initial summary writing study due to changes in their freelance work schedules.

57% of the time, even though it generated much more extractive summaries. These results suggest an intriguing gap between annotator preferences when writing and evaluating summaries.

5.1.2.3 Re-evaluating Reference-based Evaluation Metrics.

In section 5.1.1.3, we saw that the performance of automated reference-based evaluation metrics may depend heavily on the quality of reference summaries. With our newly collected freelance writer summaries, we now conduct an initial study on the effect of using higher-quality reference summaries. We focus on using Rouge-L for faithfulness evaluation on the XSUM dataset because current reference summaries are known to be highly unfaithful (Maynez et al. 2020).

In Figure 5.6, we plot the system-level Rouge-L against the human ratings for faithfulness. The left plot shows the results of computing Rouge-L with existing reference summaries from XSUM, which has a negative correlation with human ratings. This result matches our expectations because the existing reference summaries are highly unfaithful. On the right, we see the results of computing Rouge-L with the freelance writer summaries, which leads to a much more positive correlation. Hence, we see that the usefulness of reference-based evaluation is closely linked to the quality of the references, and we can improve metric correlation by using higher-quality reference summaries.

In summary, we benchmark several LLMs for news summarization across the two most popular news benchmarks. Through our experiments, we find that state-of-the-art LLMs can generate zero-shot summaries that are on par with human-written ones, with instruction tuning and RLHF being the key factors to success. We show that the existing references in these benchmarks are not suitable for summarization, and human evaluators find them to be poor summaries for the input articles. Unless the reference quality issue is addressed, comparing zero-shot, few-shot, and finetuning performance will remain an open question.

5.2 Benchmarking LLMs for Narrative Summarization.

News summarization has been the barometer for progress in the field of automatic text summarization (Radev, Hovy, et al. 2002; Rush et al. 2015; Nallapati et al. 2016; See et al. 2017; Y.-C.

Chen and Bansal 2018; L. Dong et al. 2019). In recent years, large-scale summarization datasets (Hermann et al. 2015; Narayan et al. 2018) along with pretrained language models (Devlin et al. 2019; Lewis et al. 2020; J. Zhang et al. 2020) have been the key ingredients to success, as supervised finetuning of these language models led to state-of-the-art performance for news summarization (Yang Liu and Lapata 2019; Lewis et al. 2020; J. Zhang et al. 2020; Yixin Liu, P. Liu, et al. 2022). Prior work has shown that while these models can produce incredibly fluent, human-like abstractive summaries, they tend to generate information that is inconsistent with the input article (Falke et al. 2019; Durmus, H. He, et al. 2020; Lux et al. 2020; Wilber et al. 2021). However, as we saw in §5.1, zero-shot, instruction-tuned LLMs can generate highly fluent, coherent, and faithful summaries that are on par with human-written ones.

While the progress on news summarization is impressive, prior work has shown that news articles tend to have a very particular writing structure, such as front-loading the most crucial information, and summarization systems exploit these structural cues to produce summaries (Kedzie, K. McKeown, et al. 2018; Zhong et al. 2019). Also, as we saw in §5.1, the biggest driver for the faithfulness issue of supervised news summarization systems has been the poor quality of existing references in the training data. Furthermore, learning from human feedback has been a key driver for the success of LLMs, and it's likely that these models receive higher quality signals for the summarization task (Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022). This suggests that the predictable structure of news articles and not being encumbered by poor-quality references might explain why LLMs can produce good, faithful summaries for news articles. However, this raises the question of whether LLMs are actually effective general-purpose summarizers and whether they would perform similarly well in more difficult settings where it's harder to take advantage of structural cues to produce good summaries. Therefore, in this work, we benchmark the state-of-the-art LLMs on the task of narrative summarization in order to understand whether LLMs can produce good, faithful summaries of narratives.

The goal of narrative summarization is to produce a condensed representation of the key events and characters that are central to the narrative while omitting peripheral details (Lehnert 1981).

Since narratives can have vastly differing structures, summarizers cannot simply exploit structural cues to produce good, faithful summaries – they need to understand the entire narrative and construct a good global representation of key plot points and characters along with their temporal and causal relationships (Lehnert 1981; C. Zhao et al. 2022). This makes narrative summarization an ideal benchmark for measuring progress in automatic text summarization. Prior work has explored several different types of narratives such as novel summarization (Mihalcea and Ceylan 2007; Ladhak, B. Li, et al. 2020; Kryściński, Rajani, et al. 2021), movie screenplay summarization (Papalampidi et al. 2020; C. Zhao et al. 2022), and TV series screenplay summarization (M. Chen et al. 2022; C. Zhao et al. 2022). However, on average, the inputs for these datasets are much longer than the typical token limit of existing LLMs. While recent work has explored recursive summarization as a way to mitigate this, the generated summaries tend to be lists of events rather than a coherent summary (J. Wu et al. 2021).

Therefore, in our work, to disentangle the difficulties of encoding long inputs from the challenges in understanding narratives, we instead focus on the summarization of short stories. We first collect a dataset of short stories from Reddit and generate summaries from state-of-the-art LLMs. We then ask experts to annotate these generated summaries for their faithfulness. We show that unlike with news summarization where LLMs are almost perfectly faithful, we find that the summaries generated for short stories tend to contain a lot of faithfulness errors. We then elicit faithfulness judgments from crowd-workers from Amazon Mechanical Turk for the same summaries and show that crowd-workers are unable to spot these faithfulness errors and get very low agreements with experts. Even the freelance writers and editors we hired on Upwork, albeit better than Turkers, do not get a high agreement with experts on faithfulness annotation for this task. We then explore a model-assisted annotation framework for faithfulness and show that this leads to higher agreement with expert annotators for the task. These results show that not only are LLMs not very faithful summarizers for narratives but that simply relying on crowd-sourcing to evaluate LLM outputs may be an unreliable measure of faithfulness.

5.2.1 Experimental Setup

5.2.1.1 Dataset

While there are several existing datasets for narrative summarization (Mihalcea and Ceylan 2007; Ladhak, B. Li, et al. 2020; Kryściński, Rajani, et al. 2021; Papalampidi et al. 2020; M. Chen et al. 2022; C. Zhao et al. 2022), they are ill-suited for our goals of benchmarking LLMs for narrative summarization for two main reasons. First, the average length of the input documents in some of these datasets tends to be much larger than the token limit of current LLMs. For instance, the token limit for Instruct Davinci is 4096, whereas the average length of a book chapter from Ladhak, B. Li, et al. (2020)'s dataset is 5165. Besides the issue of token limit, long narrative summarization combines challenges of narrative understanding with long document understanding and long sequence generation, which are all active areas of research on their own. Therefore, in our work, we opt for shorter narratives since we are interested in benchmarking the LLM's ability to understand and summarize narratives.

Second, and perhaps more crucially, most prior datasets for narrative summarization have sourced their target summaries from the internet. For instance, Ladhak, B. Li, et al. (2020) source their summaries from study-guide websites such as CliffNotes, M. Chen et al. (2022) source their summaries from TV recap websites such as TVMaze, and C. Zhao et al. (2022) source their summaries from websites such as Wikipedia and IMDB. Since LLMs are trained on data crawled from the internet (Brown et al. 2020; Chowdhery et al. 2022), there's a question of data contamination with existing narrative summarization benchmarks, and it's likely that the language model has seen the summaries for these narratives during training. In fact, since prior narrative summarization benchmarks contain popular books, movies, and TV series, it's possible that the LLM has seen multiple versions of summaries/recaps for these during training.

Given these two issues, we instead opt to collect a dataset of short stories from Reddit. In particular, we collect posts from two popular subreddits where community users can submit their

Dataset	Domain	Avg Doc Len
CNN/DM (Hermann et al. 2015)	News	781
XSUM (Narayan et al. 2018)	News	431
ScriptBase Papalampidi et al. 2020	Movies	24106
SUMMSCREEN (M. Chen et al. 2022)	TV Series	6612
NARRASUMM (C. Zhao et al. 2022)	Movies/TV Series	786
NovelChapters (Ladhak, B. Li, et al. 2020)	Books	5165
BookSumm (Kryściński, Rajani, et al. 2021)	Books	5102
Ours	Short Stories	850

Table 5.5: Input length for narrative summarization datasets, compared to news.

original short stories for other users to enjoy and comment on.[8] We filter our posts that are marked as NSFW and also posts that have fewer than three up-votes. This results in a collection of approximately 3000 short stories. We note that users do not write summaries for their stories, and since these stories are not popular, they're unlikely to be summarized elsewhere; therefore, there is little concern about data contamination with LLMs. Table 5.5 shows the average input length of our dataset compared to other narrative datasets and the popular news benchmarks. We see that our dataset of short stories is closer in length to news articles compared to most other narrative datasets, which tend to have much longer inputs.

5.2.1.2 Models.

We benchmark the following models for our narrative summarization:

Instruct Davinci. Ouyang et al. (2022) used instruction tuning to further improve GPT-3 and show that it leads to improved zero-shot and few-shot performance. We further show in §5.1 that this model beats the prior state-of-the-art for news summarization and produces summaries that are on par with human-written ones.

ChatGPT. This model is a successor of Instruct Davinci with a chat interface and has been further refined using RLHF (Christiano et al. 2017). Recent work has shown that this model outperforms

[8]The two subreddits we used were r/shortstories and r/shortscarystories.

its predecessor and achieves state-of-the-art zero-shot performance on a variety of NLP tasks (Bang et al. 2023).

Claude. This is the successor of the model from Yushi Bai et al. (2022) with a chat interface and has been further refined using Constituional AI (Yuntao Bai et al. 2022). This process replaces the costly process of collecting human preference for RLHF with preference labels from an AI, with human guidance coming from a set of principles for the AI to follow in making preference judgments.

We prompt all LLMs for zero-shot summarization to generate summaries for the stories. We use the following prompt for Instruct Davinci:

```
Story:  [story text]
Summarize the above story.
Summary:
```

For ChatGPT and Claude, since they are chat-based LLMs, we modify the prompt to "`Summarize the following story:  [story text]`".

5.2.2 Human Annotations.

In our initial analysis of generated summaries, we found that LLMs tend to generate fluent and coherent summaries, but tend to have very subtle faithfulness errors. Table 5.6 shows an example of a story along with a summary generated by the Instruct Davinci model. While the generated summary is fluent and coherent and does a pretty reasonable job of summarizing the events of the narrative, there's a faithfulness error highlighted in red. The summary depicts Daniel as agreeing to go to rehab after waking up for the first time in the hospital, whereas in the story, he keeps drifting in and out of sleep, and doesn't agree to go to rehab until the last time he wakes up. Given the subtleties of faithfulness errors we observed, we were unsure whether non-expert annotators would be able to reliably evaluate LLMs generated summaries. Similar concerns are echoed in

Example LLM Summary
Story: There's a beach on the Southern coast of California where the sky is pink and orange and palm trees in the view. The ocean is lightly roaring and crashing on the sand. The sound of traffic is muted by a beautiful voice. A song calling from just beyond vision's reach. Daniel searched for the source. The song continued. Daniel walked closer to the water and peered out into the deep. Then, in the moment he saw her, the sky went black and time came to a stand still; A mermaid was sitting on a rock... [TRUNCATED]
Instruct Davinci: Daniel is a young man who finds himself on a beach on the Southern California coast, where the sky is pink and orange and palm trees are in view. He hears a beautiful song coming from just beyond what he can see and goes to investigate. He finds a mermaid sitting on a rock, but the sky suddenly goes black and time stands still. He runs to the marina to find a boat so he can get to the rock, but bumps into another boat and knocks a frozen person overboard. He jumps in to rescue them and eventually finds himself on the rock with the mermaid, but she has suddenly changed from a beautiful creature into a gray seal. He panics and crashes the boat, knocking himself unconscious. When he wakes up he is in a hospital and his parents are discussing sending him to rehab. **Daniel agrees, and then falls back asleep**. He wakes up again on the beach and the seal is there, singing the same song. He jumps into the water and the seal bites him, but his skin breaks the teeth. Daniel smiles and brings them back to the beach, and when he wakes up in the hospital again he tells his father rehab is a good idea.

Table 5.6: Example of a summary generated by Instruct Davinci. The overall summary is very fluent and mostly correct, except for a subtle error shown in **red**.

recent work, which shows that existing human evaluation approaches are not very robust (Yixin Liu, Alexander R. Fabbri, et al. 2022).

Given these concerns, we recruit three NLP experts who are working on text summarization and collect a set of gold standard labels for faithfulness evaluation. In particular, we ask the annotators to annotate a set of 33 LLM summaries for faithfulness, followed by an adjudication round to resolve any disagreements in the annotation. Table 5.7 shows the percentage of generated summaries considered faithful according to expert annotators for each of the LLMs. We observe that a majority ($60 - 70\%$) of LLM-generated summaries contain faithfulness errors. This is in stark contrast to prior work on news summarization, which showed that LLMs are almost perfectly faithful when generating zero-shot summaries for news articles (see §5.1; Goyal, J. J. Li, and Greg Durrett 2022). This suggests that LLMs may not be reliable general-purpose summarizers beyond news summarization.

We then collect annotations for faithfulness from non-experts using crowd-sourcing via Amazon Mechanical Turk (MTurk). We also enlist freelance writers and editors from Upwork since they tend to be more reliable than Turkers. Figure 5.7 shows the task description shown to the annotators for the annotation task, and Figure 5.8 shows the detailed instructions along with an example of how to annotate for faithfulness errors. In some instances, LLMs can add some additional commentary to the end of the summary, and for this task, we ask annotators to ignore the commentary. We ask the crowdworkers to annotate the same set of 33 summaries that were annotated by experts, and we elicit 3 independent annotations per summary. We take the majority label for each summary as the faithfulness annotation for the summary. We are interested in understanding how well the crowdworker consensus matches up with expert annotators for faithfulness evaluation.

Table 5.7 shows the percentage of summaries considered faithful according to non-experts. We see that annotators on MTurk generally find LLMs to be relatively faithful and only find $4-15\%$ of the summaries to be unfaithful, depending on the LLM. In particular, the most effective LLM, according to Turkers (ChatGPT), is actually the one deemed least faithful according to expert annotators. While the annotators on Upwork find LLM-generated summaries less reliable than Turkers, they still annotate a larger percentage of summaries as faithful compared to expert annotators (70% vs. 33%, respectively). Overall, we find that non-experts really struggle to identify subtle faithfulness errors when summaries are generally fluent and coherent. This suggests that standard human evaluation on crowd-sourcing platforms may not be a reliable way to evaluate models as they become increasingly better.

5.2.3 Model-Assisted Human Evaluation.

Given the difficulties that non-experts face in identifying faithfulness errors for narrative summarization, we investigate whether an AI model can help improve the quality of crowd-sourced annotations. Specifically, we examine whether providing model-generated critiques and comments can lead crowd workers to agree more with expert annotators. Prior work has shown that LLM critiques can help annotators identify errors they may otherwise miss (Saunders et al. 2022). We

Model	Experts	MTurk	Upwork
Instruct Davinci	0.36	0.88	0.70
ChatGPT	0.27	0.96	0.68
Claude	0.36	0.85	0.73

Table 5.7: Faithfulness scores assigned by experts vs crowd workers for LLMs on our narrative summarization dataset. We find that even for the best LLMs, around 70% of the generated summaries, contain faithfulness errors, according to experts, however, non-experts struggle to identify these errors.

adopt a similar approach and ask Claude to pinpoint any inconsistent details between a summary and the source text. We then provide these critiques to crowd workers when asking them to assess how faithful the summary is to the original text. Table 5.8 shows Claude's critique for the summary in Table 5.6. While Claude incorrectly concludes that the summary is faithful overall, it does correctly identify a subtle yet important faithfulness error. We hypothesize that even imperfect model critiques can make crowd workers more attentive to potential issues, helping them better identify faithfulness errors.

To further evaluate the reliability of the crowd-sourced annotations, we conduct an additional study using a second group of freelance writers and editors recruited through Upwork, and ask them to perform the faithfulness annotation on the same set of summaries. This new group is provided with Claude's faithfulness critiques for each summary but is explicitly informed that the model's comments may contain errors and that they should use their own best judgment. Figure 5.9 shows the modified task description that is shown to this group of annotators. We then quantitatively compare the level of agreement with expert annotators between the two groups - those who received Claude's potentially flawed critiques versus those who did not. Interestingly, the group exposed to Claude's critiques exhibits markedly higher raw agreement with the experts (58% vs. 45% for the unexposed group). In terms of Cohen's Kappa, the 58% agreement translates to a kappa score of 0.22 for the group with critiques (vs. 0.12 without critiques), which can be interpreted as fair agreement according to the guidelines provided by Landis and Koch (1977). This indicates that

Example LLM Critique
Here are the key details summarized: Daniel finds a mermaid on a rock who turns into a gray seal. He crashes his boat trying to get to the rock and knocks himself unconscious. He wakes up in a hospital and agrees to go to rehab. He has another dream where the seal bites him but his skin breaks the seal's teeth. He wakes up again and agrees rehab is a good idea. The only inconsistent detail I identify is that the summary says Daniel agrees to go to rehab after waking up in the hospital the first time, but in the story he actually agrees after waking up in the hospital the second time. Other than that, the summary is consistent with the key events and details in the story. The summary may be missing some details but it accurately captures the core parts of Daniel's experiences, dreams, and decision to go to rehab. So overall, aside from the minor inconsistency regarding when Daniel agrees to rehab, the summary is consistent with the story.

Table 5.8: LLM generated critique for the example summary shown in Table 5.6. Even though the model arrives at the wrong conclusion for overall faithfulness, it is correctly able to identify the faithfulness issue in the summary, shown in **green**.

presenting crowd workers with AI-generated critiques, even imperfect ones, can improve their ability to correctly identify faithfulness issues in the summaries. The resulting crowd-sourced annotations are substantially closer to expert judgments. We note, however, that a 58% agreement rate (0.22 kappa score) with experts is still far from ideal. This preliminary approach of simply showing Claude's raw critiques to workers does not yet provide a sufficiently reliable mechanism to completely replace expert evaluation. Further iterations on how to best leverage AI assistance in the crowd-sourcing process is needed.

5.2.4 Qualitative Analysis of Narrative Summaries

While the summaries generated by LLMs for the narratives tend to appear really good, they usually 1-2 errors that distort key points of the narrative. For example Table 5.9 shows an example of a summary generated by ChatGPT. While at the surface level the summary looks good, the part highlighted in red is incorrect and distorts a pivotal part of the narrative. In the story, the team

Example LLM Summary
ChatGPT: A rescue team goes into a narrow cave to find a participant who had fallen down a blocked-off section. They drop a radio down to him and start a multiple day rescue operation, during which they take turns speaking with the man who remains surprisingly calm. When they finally extract him, he apologizes for losing their radio, but before they can say anything, the radio flares to life with the same voice asking when they will come to get him. The team **realizes that the man they rescued was not the one who had fallen down the shaft**, and they wonder who they had actually been talking to.

Table 5.9: Example of a summary generated by ChatGPT. The overall summary is very fluent and mostly correct, except for the error shown in **red**, which changes a key part of the story.

realizes that the man they rescued was the one who had actually fallen down the shaft, but he wasn't the person they had been talking to on the radio. Table 5.10 shows another example, where all three LLMs make the same error. Their summaries portray a pleasant reminiscing conversation between the narrator and mother about the past. In truth, the original exchange was unpleasant and implied a history of abuse by the mother. Though easy to miss by non-experts, these types of factual distortions misrepresent critical narrative elements, making the summary unfaithful to the source. While superficially smooth, closer inspection reveals that the LLMs often alter key details in their summaries.

While a majority of the summaries contain faithfulness errors, these models do exhibit strengths that can be built upon in an AI-assisted summary writing system. Most notably, the summaries generated tend to have good structural coherence and identify the main themes and narrative arcs well, even if some of the details are incorrect. This provides a strong skeleton that humans can then flesh out and correct to build good summaries. In addition, as evidenced by Claude's ability to provide high-level commentary on its own summaries in Table 5.11, LLMs show promising capabilities for abstractive summarization. The commentary generated by Claude (shown in green) effectively captures the overall point or meaning of the narrative, correctly identifying the main goals, tensions, and resolutions even when the summary contains faithfulness errors (shown in red). This demonstrates that LLMs are capable of higher-order comprehension of themes, motives, and general messaging. These strengths of structural coherence and high-level abstraction suggest

Example LLM Summaries
Story: [TRUNCATED] ... While standing there, I feel numb. My fingers grip the bouquet tight. I remember Mom, a sweet woman. Her smile, her strong hands, and witty humor . . . After several minutes, I gather the courage to speak to her. "How's it going, mom?" I feel strong emotions coming. Mom's voice answers me in return: You brat! I took care of you and your brother! I gave you your own closet. But you scratched it like a cat! You! You didn't know how to wash in the tub, so I showed you how! I laugh at the whisper. "You're telling a different story, mom." I smile. "At least you won't hurt me or Hector anymore."
Instruct Davinci: The narrator is visiting a cemetery, and has a strange gift where they can hear the voices of the dead. As they explore, they come across several graves that are speaking to them, including one belonging to a baby, a man who was prejudiced against people of color, and a beloved mother. The narrator talks to their mother's grave and **reminisces, and finds that even in death, their mother still has plenty of wit and humor.**
ChatGPT: The protagonist hates going to cemeteries because they have the ability to hear the dead and the dead can hear them too. They walk through the cemetery and hear various voices and stories from the graves. They eventually make it to their mother's grave and have a conversation with her, **reminiscing about their past. The story ends with the protagonist feeling comforted by their mother's voice.**
Claude: The narrator has the unusual ability to hear the voices of the dead when visiting cemeteries. The narrator finds cemeteries unpleasant to visit because of this ability. While walking through a cemetery and hearing various voices from graves, the narrator arrives at the grave of the narrator's mother, marked by an angel statue. The narrator speaks to the mother's spirit and **reminisces about memories of her. Despite the strange ability, the narrator finds comfort communicating with the mother's spirit.** The story suggests the narrator has a complex relationship with the supernatural ability to hear the dead.

Table 5.10: Summaries for one of the stories in our dataset. We see that all three LLMs make the same error in understanding a key part of the narrative.

LLMs can serve as AI assistants that provide a first draft for humans to refine and fact-check. The human can verify and fill in details while benefiting from the summarization and high-level analysis provided by the LLM. This allows humans to focus their efforts on where they add the most value - nuanced reasoning and fact-checking. Such a collaborative system would combine the strengths of both humans and AI. Overall, while work remains to improve faithfulness, LLMs exhibit abilities in abstraction and coherence that can pave the way for hybrid AI-assisted summarization.

Example LLM Summary
Claude: The goddess Aphrodite is upset that a poll shows Athena, the goddess of war, is considered fairer than her. This threatens Aphrodite's power as the goddess of love and beauty, which comes from worship and interest in her domain. Aphrodite asks her son Cupid to shoot all of Earth with an arrow of love directed at her to fix the problem, but Cupid says he does not have the power to do that and it would not be effective. Aphrodite then decides to travel to the mortal realm herself for the first time in a long while to better understand how human attraction and flirting work nowadays. After being turned down by a gay man, she meets another man who gives her a cell phone after a one night stand. **The story suggests that Aphrodite must learn how modern human attraction and relationships work if she wants to maintain her power as the goddess of love.**
Claude: Destiny goes to visit her best friend Heaven at her house. Heaven's father lets Destiny into the house and leads her down to the basement, claiming Heaven has been playing down there all week. **Once Destiny reaches the bottom of the stairs**, she discovers Heaven's dead body in the dark basement. Heaven's ghost then whispers to Destiny that now they can play together in the basement. **The story suggests that Heaven's father was responsible for her death and lured Destiny into the basement to harm or kill her as well.**

Table 5.11: Example summaries showing Claude's ability to generate commentary in addition to the story (highlighted in **green**).

5.3 Related Work

News summarization has been the key task that has been used to measure progress in the field of automatic summarization (Radev, Hovy, et al. 2002; Rush et al. 2015; Nallapati et al. 2016; See et al. 2017; Y.-C. Chen and Bansal 2018; L. Dong et al. 2019). In recent years, supervised finetuning of language models with large-scale news summarization datasets has emerged as the main paradigm for building state-of-the-art summarization systems (Yang Liu and Lapata 2019; Lewis et al. 2020; J. Zhang et al. 2020; Yixin Liu, P. Liu, et al. 2022). While these systems tend to produce highly fluent, human-like summaries, recent work has shown that they suffer from generating information that is inconsistent with the source article (Falke et al. 2019; Durmus, H. He, et al. 2020; Lux et al. 2020; Wilber et al. 2021). As such, there has been increased attention in the literature towards improving faithfulness of summarization systems, as it is an essential requirement in order to deploy these systems in real-world scenarios (see chapter 4; Matsumaru et al. 2020;

Z. Zhao et al. 2020; Y. Dong, S. Wang, et al. 2020; Goyal and Greg Durrett 2021; Xu et al. 2020; S. Chen et al. 2021; Zhu et al. 2021). In our work, we benchmark LLMs for news summarization and show that instruct-tuned LLMs outperform the prior state-of-the-art and perform on par with humans. We also show that faithfulness in news summarization is no longer a major concern with these LLMs and that faithfulness issues plaguing prior models were mainly stemming from noisy references in the training data, as also alluded in prior work (Maynez et al. 2020; Kang and T. B. Hashimoto 2020).

In addition to all the work at improving news summarization systems, there is also a large body of work looking at improving the evaluation of summarization; many automated evaluation metrics have been proposed over the years (Lin 2004; Papineni et al. 2002; T. Zhang, Kishore, et al. 2019; Sellam et al. 2020; Durmus, H. He, et al. 2020; Kryscinski, McCann, et al. 2020). Recent work, however has shown that these evaluation metrics do not correlate well with human judgments (see §2.1; Durmus, H. He, et al. 2020; Alexander R Fabbri et al. 2020; Durmus, Ladhak, et al. 2022). In this work, we evaluate the effectiveness of automatic metrics for evaluating LLMs and show that the usefulness of reference-based evaluation is closely linked to the quality of the references, which is consistent with prior work in machine translation (Freitag et al. 2020). In addition, we show that improved references can lead to improved correlation with human evaluation of summarization.

While the majority of the work in improving and evaluating summarization systems has focused on news summarization, there is a significant body of work that looks at summarizing documents beyond news articles, such as scientific articles (Cohan et al. 2018; Yasunaga et al. 2019), dialogues (Gliwa et al. 2019; Y. Chen et al. 2021), novels (Mihalcea and Ceylan 2007; Ladhak, B. Li, et al. 2020; Kryściński, Rajani, et al. 2021), and screenplays (Papalampidi et al. 2020; C. Zhao et al. 2022; M. Chen et al. 2022; C. Zhao et al. 2022). These tasks aim to understand whether automated summarization systems can understand and summarize inputs with broader narrative and discourse structures than typically seen in news articles. Similarly, in our work, we aim to understand whether the recent success of LLMs in zero-shot news summarization transfers to narrative summarization. While prior work on narrative summarization has focused on chapter-level (Ladhak, B. Li, et al.

2020; Kryściński, Rajani, et al. 2021) and book-level (Mihalcea and Ceylan 2007; Kryściński, Rajani, et al. 2021) summarization of novels, our work instead chooses to focus on short stories from Reddit. There are several reasons for this decision: 1) novels typically contain input lengths that are much longer than LLM context windows 2) summaries for popular novels are readily available on the web, which contaminates the evaluation, and 3) we did not want to conflate challenges with processing long context with the LLM's ability to understand and summarize narratives.

Recent years have seen an increase in human-AI collaboration, particularly with using AI to augment human decision-making (Lai et al. 2021) in high-stakes domains such as medical diagnoses (Lakkaraju et al. 2016) and credit risk prediction (Chromik et al. 2021). Several recent works have shown that AI can augment the human ability to complete natural language tasks more accurately. Fan et al. (2020) use AI to generate natural language briefs that assist humans in fact-checking claims both faster and more accurately. Gehrmann, Strobelt, et al. (2019) use language models to help improve human accuracy in detecting AI-generated text from human-written text. Most closely related to our work is the work by Saunders et al. (2022), which showed that using model-generated critiques can help humans identify errors in generated summaries that they might otherwise miss. Similarly, in our work, we show that model critiques can help humans identify faithfulness errors that they might otherwise miss.

5.4 Conclusion and limitations

In §5.1, we conducted a comprehensive human evaluation of 10 LLMs across the two most popular news summarization benchmarks. Through our experiments, we find that the state-of-the-art LLM performs on par with summaries written by freelance writers, with instruction tuning being the key factor for success. Beyond these findings, our work highlights the crucial role of good reference summaries in both summarization model development and evaluation. Unless the reference quality issue is addressed, comparing zero-shot, few-shot, and finetuning performance will remain an open question, and the current benchmarks will provide limited value when used with reference-based evaluation. Even when we address the quality issue and conduct a human evaluation with high-quality references, we observe a significant amount of individual variation from annotators. Due to these factors, evaluations for single document news summarization may be reaching their limits. The same, however, cannot be said for the summarization in general – as we show in §5.2, LLMs still struggle to produce faithful summaries of narratives, unlike news articles. This suggests that while instruction-tuning may have improved summarization capability for news, it does not lead to improved summarization in general. A possible explanation is that news summarization is typically included as one of the tasks during instruction tuning Ouyang et al. 2022.

We believe that there is much research beyond our benchmarking effort that needs to be done to better understand the effect of instruction tuning. Here, we hypothesize three aspects that could account for the success of instruction tuning. First, the quality of the summarization data used in instruction tuning can serve an important role. Our findings in §5.1.1 show that currently, we are finetuning language models on low-quality training data, which can account for their ineffectiveness. At this point, we cannot rule out the possibility that when finetuned on a higher-quality dataset, finetuned LMs may perform much better. Second, the learning algorithm used for instruction tuning can be important (Ouyang et al. 2022). While the exact training details are unknown, the success of Instruct Davinci might be credited to "learning from human feedback" (LHF; Stiennon et al. 2020; Ziegler et al. 2019). Contrary to supervised finetuning that trains systems on written summaries, learning from human feedback trains systems from binary labels of human preferences. As we observe in §5.1.2.2, there is a discrepancy in how annotators write and rate summaries. While it is possible that LHF has merits over the supervised learning/finetuning approach in exploiting this discrepancy, more analysis is needed to validate this hypothesis. Third, multi-task learning could be an important factor in the success of LLMs. Instruct Davinci is trained on a diverse distribution of inputs, and many previous studies have confirmed the effectiveness of multi-task learning. We need need further work to understand how summarization benefits from learning on other tasks and what tasks are helpful vs. harmful for summarization performance. Furthermore, we need to understand whether instruction tuning on more difficult summarization tasks, such as narrative summarization, can lead to improved general summarization capability.

Our work also highlights the difficulties in evaluating high-performing LLMs. Recent work advocates for using fine-grained semantic units to match with reference summaries Yixin Liu, Alexander R. Fabbri, et al. (2022) in order to get a more reliable human evaluation. However, as our evaluation points out, not only are the existing reference summaries unreliable, but the summaries written by well-paid freelance writers also may not significantly outperform LLM summaries. Therefore, defining reference summaries as the ground truth may be overly restrictive as LLMs are approaching or even exceeding average human-level performance. Furthermore, we show that simply asking crowd-workers to evaluate LLM outputs may no longer be reliable as LLMs become increasingly fluent and coherent. We note a large gap in the ability of crowd workers to identify deficiencies in LLM-generated summaries compared to experts for narrative summarization. However, relying on expert evaluations is not scalable, and we show that investing in AI-assisted human evaluations may offer a viable avenue to scaling up evaluation. For instance, one possible avenue would be to allow evaluators to interact with models through clarification dialogues to elicit information to make more informed judgments. We leave this exploration for future work.

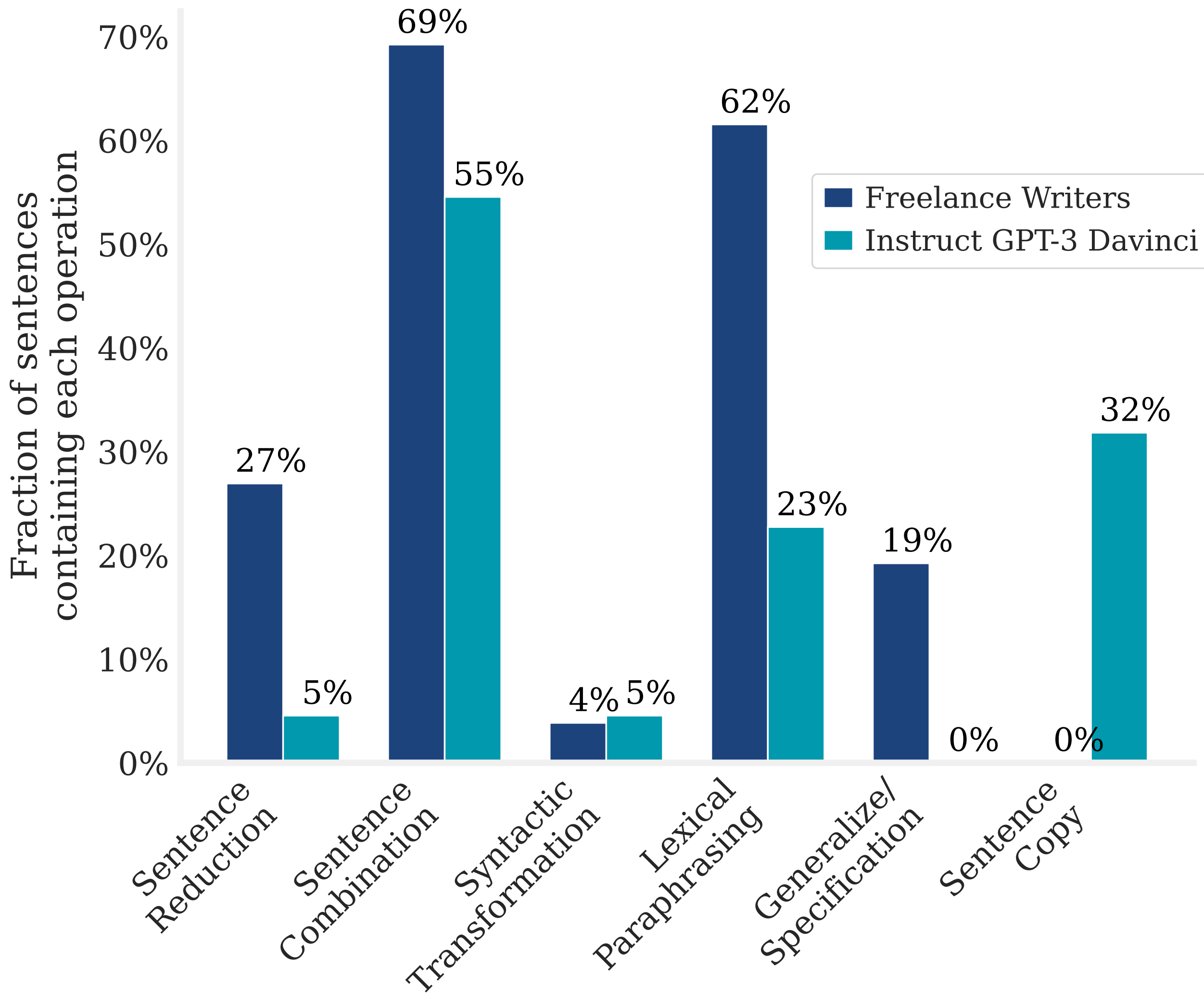


Figure 5.4: Distribution of cut and paste operations in the summaries written by freelance writers and by Instruct Davinci. By comparison, human written summaries contain more lexical paraphrasing and sentence reduction, whereas the Instruct Davinci model has more direct copying from the article.

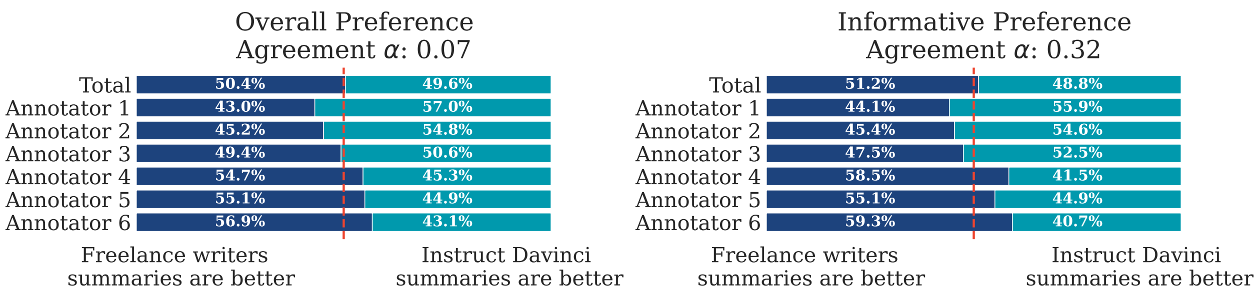


Figure 5.5: Human evaluation results comparing summaries written by freelance writers and summaries generated by Instruct GPT-3 Davinci. On aggregate, annotators equally prefer freelance writers and Instruct Davinci. However, there is high variability in individual annotators' preferences. Notably, annotator, 1 writes abstractive summaries but prefers the more extractive Instruct Davinci summaries.

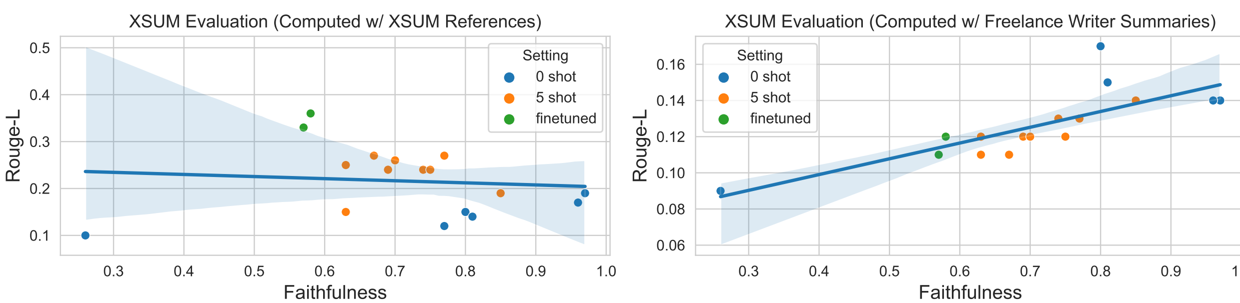
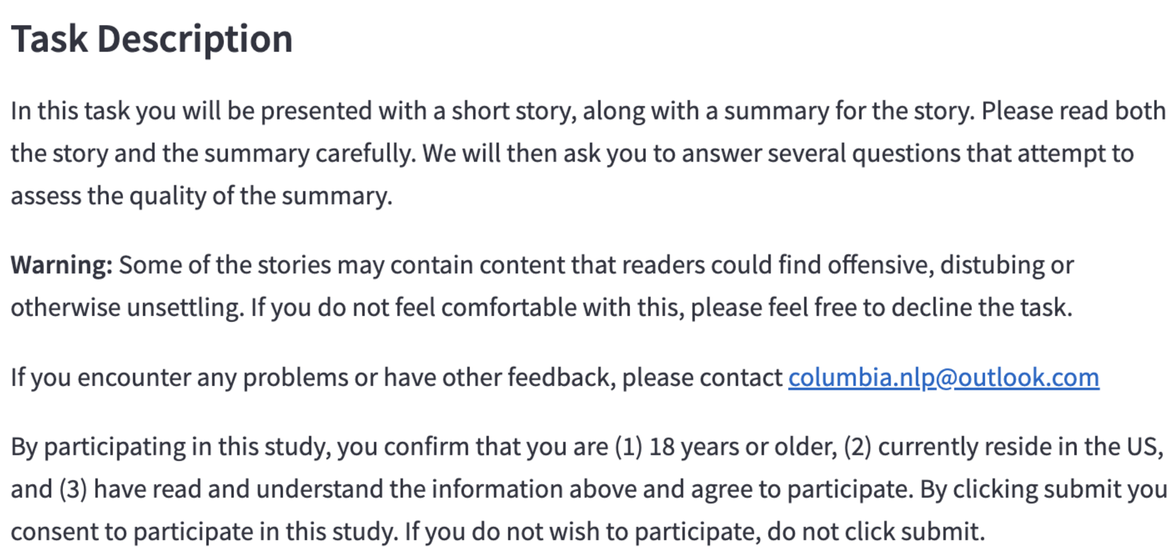


Figure 5.6: System-level Rouge-L vs. annotator rating of faithfulness. The left plot is computed with XSUM references, where the correlation is weak, and the right plot is computed with the freelance writer summaries, where the correlation is much improved.



Task Description

In this task you will be presented with a short story, along with a summary for the story. Please read both the story and the summary carefully. We will then ask you to answer several questions that attempt to assess the quality of the summary.

Warning: Some of the stories may contain content that readers could find offensive, distubing or otherwise unsettling. If you do not feel comfortable with this, please feel free to decline the task.

If you encounter any problems or have other feedback, please contact columbia.nlp@outlook.com

By participating in this study, you confirm that you are (1) 18 years or older, (2) currently reside in the US, and (3) have read and understand the information above and agree to participate. By clicking submit you consent to participate in this study. If you do not wish to participate, do not click submit.

Figure 5.7: Task description shown to the crowdworkers for the annotation task.

Detailed Instuctions with Example

Please read the story carefully and try to understand the points that are central to the plot.

Story

> They shot the six cabinet ministers at half-past six in the morning against the wall of a hospital. There were pools of water in the courtyard. There were wet dead leaves on the paving of the courtyard. It rained hard. All the shutters of the hospital were nailed shut. One of the ministers was sick with typhoid. Two soldiers carried him downstairs and out into the rain. They tried to hold him up against the wall but he sat down in a puddle of water. The other five stood very quietly against the wall. Finally the officer told the soldiers it was no good trying to make him stand up. When they fired the first volley he was sitting down in the water with his head on his knees.

Below are two example summaries for this story:

Summary A

> The short story describes the execution of six cabinet ministers by firing squad early in the morning during heavy rain. One of the ministers is sick with typhoid fever and too weak to stand, so he is brought downstairs and shot while sitting in a puddle of water with his head on his knees. **The story paints a vivid image of human suffering.**

Summary B

> Six cabinet ministers are executed by firing squad early in the morning during heavy rain. One of the ministers is sick with typhoid fever and too weak to stand, so he is shot while **lying in a puddle of water** near the hospital wall.

Question: Is the information in the summary consistent with the story?

The first question asks you to judge whether or not the summary is consistent with the original story. In this case you want to make sure the events and details described in the summary are mentioned in the actual story. For example, *Summary B* is **not** consistent with the story as the minister was sitting in the puddle of water and not lying in it. Note that *Summary A* is consistent with the story. The last bold sentence in *Summary A* contains some commentary on the story, and we ask you to ignore commentary in evaluating consistency.

Figure 5.8: Detailed instructions for the task along with an example of how to annotate for faithfulness.

Task Description

In this task you will be presented with a short story, along with a summary for the story. Please read both the story and the summary carefully. We want you to identify whether or not the summary contains any details that are inconsistent with the original story. We also provide the ouput of an AI agent that attemps to point out some issues with the summary. The AI agent's output is provided to help you in assessing the summary, but ultimately you should rely on your own judgment. The AI agent may not always be correct in its assessment.

Warning: Some of the stories may contain content that readers could find offensive, distubing or otherwise unsettling. If you do not feel comfortable with this, please feel free to decline the task.

If you encounter any problems or have other feedback, please contact columbia.nlp@outlook.com

By participating in this study, you confirm that you are (1) 18 years or older, (2) currently reside in the US, and (3) have read and understand the information above and agree to participate. By clicking submit you consent to participate in this study. If you do not wish to participate, do not click submit.

Figure 5.9: Task description shown to the crowd workers who receive AI assistance. They are told that the critiques may not be accurate, and are instructed to ultimately rely on their own judgments.

Conclusion and limitations

In this thesis, we explore the problem of faithfulness in abstractive summarization. In chapter 2, we make significant contributions to advancing faithfulness evaluation in summarization. We demonstrate substantial flaws with existing evaluation approaches, including both automated metrics and human evaluation. To address the deficiencies of current automated metrics, we develop a novel method for training reference-free evaluation metrics that are more robust to distributional shifts. Our proposed approach leads to a metric that outperforms prior work at ranking state-of-the-art summarization systems. Similarly, to improve upon current approaches to human evaluation, we propose a new evaluation framework that enables a more holistic assessment of faithfulness by accounting for the abstractiveness of systems. Through advances in both automated and human evaluation, this chapter enables a more rigorous and meaningful evaluation of faithfulness in summarization.

In chapter 3, we make key contributions towards explaining the cause of faithfulness issues in modern summarization systems. First, we propose a novel contrast-based method for error attribution that dramatically outperforms prior work at identifying the origins of hallucination in generated summaries. Moreover, we demonstrate the broad applicability of our approach by incorporating its ideas into an existing method, TracIn, yielding substantial improvements. Second, we reveal the effect of distributional biases in pretrained language models, showing that strong associations learned during pretraining can cause summarization models to hallucinate information in a biased manner. We further explore mitigation strategies during finetuning, demonstrating that while they can reduce overall hallucination rates, the remaining hallucinations still closely reflect intrinsic biases in the pretrained models. Through advanced error attribution and analysis of pretrained model biases, this chapter provides critical insights into the roots of faithfulness issues in modern abstractive summarization.

Building on the findings from previous chapters, chapter 4 develops effective methods for improving faithfulness in summarization systems. First, we propose a novel approach that adapts the level of abstractiveness for each input to enhance overall faithfulness. Our method yields systems that are both more faithful and more abstractive compared to baselines. Second, we leverage our proposed error attribution technique to clean existing datasets, enabling the training of models with reduced hallucination. Models trained on datasets cleaned with our approach generate markedly fewer hallucinations than both baseline systems and models trained using other data cleaning techniques. Through adaptive abstractiveness and targeted data cleaning, this chapter provides impactful techniques for mitigating faithfulness issues in state-of-the-art abstractive summarization.

Finally, chapter 5 focuses on characterizing the summarization capabilities of LLMs and assessing the extent of the faithfulness problem for these models. First, we demonstrate that instruction-tuning and RLHF are key for enabling LLMs to achieve high-quality zero-shot summarization in the news domain, with state-of-the-art LLMs generating summaries comparable to human-written ones. However, this capability does not extend to narrative summarization, where even advanced LLMs struggle to consistently produce faithful summaries. Finally, we highlight the difficulty in evaluating high-performing LLMs, showing that simply crowdsourcing evaluations of LLM outputs may no longer be reliable as fluency and coherence improve. We observe a substantial gap between crowd workers and experts in identifying deficiencies in LLM-generated narrative summaries. Through rigorous analysis, this chapter provides important insights into the summarization capabilities and limitations of LLMs.

5.5 Takeaways

Current benchmark datasets, especially for news summarization, are inadequate. Rather than carefully constructing datasets grounded in real-world use cases, prior benchmarks have relied on the ability to source large-scale datasets from the web using simple heuristics to create target summaries (Hermann et al. 2015; Rush et al. 2015; Narayan et al. 2018). While performance on these datasets has steadily improved over the years, this progress has not translated into improved summarization capabilities. This is unsurprising, as our work shows humans do not find these references to be good summaries. As we show in subsubsection 5.1.1.3, our human evaluators find only 37% of the reference summaries in the XSum dataset are faithful – optimizing for these references means building a summarizer that frequently hallucinates information. Indeed, our work in §3.1 shows how introducing a small amount of noise (perturbations) into the training data can result in hallucinations in generated outputs. We propose a novel contrast-based attribution method that can be used to identify noisy instances in the dataset that can lead to undesirable generated outputs. In §4.2, we use our error attribution method to clean existing datasets and show that removing noisy instances from training leads to reduced hallucinations in the generated output. Even incorporating a few of these noisy instances as in-context examples leads to a dramatic drop in the summarization performance of LLMs, as we show in §5.1. We should stop using these flawed datasets as benchmarks for summarization capability and work towards carefully constructed datasets grounded in real-world use cases to make meaningful progress on summarization.

Improving faithfulness evaluation metrics requires improving evaluation benchmarks. Recent advancements in abstractive text summarization have primarily been evaluated using ROUGE, which fails to capture critical summarization aspects like faithfulness. While recent work has proposed new faithfulness metrics with remarkably improved performance, our analysis in §2.1 indicates that these improvements stem primarily from flaws in benchmark construction rather than actual improvements in evaluating faithfulness. Specifically, current benchmarks allow metrics to exploit simple spurious correlations – when we control for these spurious measures, we see a dramatic drop in performance. We propose a method to reduce reliance on spurious correlations and demonstrate improved performance on existing benchmarks. In particular, we show that our metric is more robust to distribution shits and does remarkably better than prior work at ranking systems that are close to the state-of-the-art. The benchmarks themselves, however, may not represent faithfulness errors of current state-of-the-art systems, as they contain summaries from systems trained on noisy datasets like XSum and CNN/DM. The resulting errors likely differ from systems not trained on these datasets. Indeed, in §5.1.1.3, we find that top-performing metrics on existing benchmarks achieve poor system-level correlation with human scores when ranking zero-shot and few-shot LLM summarizers. To build better faithfulness evaluation metrics, we must first design better benchmarks containing more realistic errors made by current state-of-the-art LLMs. Furthermore, constructing benchmarks using models exposed to high-quality supervision would better represent the types of errors we need to detect. With improved benchmarks, we can develop metrics that reliably identify salient faithfulness issues. Progress in both benchmark design and metric development is key to properly evaluating summarization faithfulness.

Simply performing human evaluation does not guarantee reliable comparisons. Recent work has increasingly relied on human evaluation as a more reliable measure of summarization capability. However, simply performing human evaluation does not necessarily lead to a reliable measure of summarization performance. As discussed in §2.2, naively comparing human faithfulness scores of summarization systems can lead to misleading conclusions, potentially at the cost of improving abstraction. While we propose an improved framework that can more holistically measure faithfulness and abstraction, this approach still depends on reliable human evaluation of the generated summaries. As we show in § 5.2, this cannot be taken for granted with LLM-generated summaries. We find that crowdworkers struggle to accurately identify faithfulness errors in LLM-generated summaries for narratives. Furthermore, in §5.1.2, when comparing LLM summaries to human-written ones with qualitatively better abstraction, we show that it is hard to get high inter-annotator agreement and there's a large variability in individual stylistic preferences. While we show some initial promising results using AI-assisted human evaluation (§5.2) to alleviate these issues, substantial work is needed to develop robust human evaluation protocols to obtain reliable assessments, especially as the capacities of LLMs continue improving.

5.6 Limitations

While this dissertation makes several contributions toward advancing the understanding and improvement of faithfulness in abstractive summarization systems, there are some overarching limitations that provide opportunities for future work.

One fundamental limitation is that most of the analysis focuses on a single domain – news summarization. While news articles are a popular choice for studying summarization systems, they have relatively predictable discourse structures compared to other domains like conversational dialogue or narratives. An interesting direction would be to investigate whether the findings around faithfulness evaluation, error attribution, and mitigation strategies hold up when applied to other domains beyond news. Expanding the scope could reveal new challenges and opportunities. For instance, our preliminary work on narrative summarization suggests that LLM-generated summaries for narratives introduce more subtle inaccuracies that are difficult for non-experts to detect, unlike news summarization. Further characterization of the types of faithfulness errors arising in narrative summarization could enable more tailored detection and mitigation techniques. More broadly, expanding the range of summarization tasks studied would likely uncover new error modalities and test the limitations of current evaluation methods. This could drive further research into more robust frameworks for faithfulness assessment, attribution, and improvement that generalize across different genres. Tackling a more comprehensive range of summarization tasks would ultimately lead to more versatile techniques and substantial progress in developing truly faithful summarization systems.

Second, our work examines faithfulness in the context of summarizing individual, relatively short documents. However, summarizing longer documents or multiple documents raises additional challenges for faithfulness that we do not address. Summarizing a long document like a book involves more complexity in ensuring that a concise summary accurately captures the full scope of concepts, relationships, and development arc within the source text. Maintaining faithfulness for multi-document summarization poses difficulties in reconciling contradictions across sources and representing different perspectives accurately. Generating summaries of longer or multiple documents likely requires a more nuanced approach to evaluating faithfulness. New methods may be needed to assess faithfulness for capturing the overarching narrative or themes in a long text. Similarly, new techniques could help check that a multi-document summary faithfully represents the diversity of viewpoints and inconsistencies across different sources. Our current work focuses solely on individual short documents, but expanding the study of faithfulness to longer and multi-document summarization remains an important direction for future work. Open research questions exist around defining and measuring faithfulness in these more complex summarization tasks.

Third, our work focuses on faithfulness, which aims to ensure that the information in the generated summary is consistent with the input document. While faithfulness is an important objective for many summarization tasks, it has limitations. Faithfulness only verifies that the facts in the summary match the facts in the source document but does not evaluate whether the summary provides sufficient context or background information to enable understanding. For instance, we may want summarization systems that can augment an input document with relevant contextual details that are factually accurate, even if not present in the original text. Providing such factual background information raises new challenges beyond faithfulness. The system must synthesize and integrate contextual details that enrich comprehension of the summary rather than simply reproducing content from the input. Furthermore, the system must ensure the accuracy of synthesized background facts, which requires additional capabilities beyond verifying consistency with an input text. Developing summarization systems that can provide factual background details beyond what is stated in an input document may necessitate new methods to fact-check, verify, and validate generated content. Overall, faithfulness provides a starting point, but richer summarization requires capabilities to synthesize and validate new factual information that provides helpful context beyond what is stated in the original document.

Finally, all our experiments and analyses have focused solely on the English language, and it remains to be seen to what extent the findings will transfer to other languages. Comparative studies across languages can shed light on the language-specific intricacies involved in abstractive summarization. For instance, the syntactic and morphological complexity of a language could impact the difficulty of summary generation. Languages with freer word order and rich morphology like Finnish and Turkish may present distinct challenges not present in English summarization, and could therefore have different types of faithfulness errors. Further cross-lingual research is needed to determine the extent to which our conclusions generalize and to reveal language-specific summarization issues.

In summary, this dissertation makes valuable contributions to advancing faithfulness in abstractive summarization but has limitations in scope that provide opportunities for future work. Broader investigations across genres, tasks, and languages will be important to gain a comprehensive understanding of faithfulness in abstractive summarization more generally. Nonetheless, the dissertation provides a strong foundation and valuable insights to build upon in future work.

5.7 Future Work

While we have highlighted several fruitful avenues for future work in the limitations section, we provide some additional promising directions that could further advance this research area.

Current benchmarks for evaluating summarization models have some critical shortcomings that need to be addressed in future work. Many existing datasets rely on web-scraped data and use heuristics to generate target summaries. This process inevitably introduces noise and a lack of quality control into the datasets, leading to issues with summary faithfulness. As the field shifts towards leveraging LLMs for summarization, using datasets with loosely curated web-sourced summaries also raises the risk of test set contamination, resulting in inflated metrics that do not accurately measure progress. Going forward, the summarization community should focus efforts on constructing new datasets that better represent real-world applications and use cases. With the representational power of LLMs, researchers need not depend on massive datasets of questionable quality. Smaller, carefully curated benchmarks that evaluate precise capabilities and model behaviors will be more insightful. Potential directions include datasets that test summarizing longer documents, summarizing complex reasoning, summarizing opinions, summarizing dialogue, and summarizing collections of documents. The end goal should be benchmarks that encourage building summarization systems that are useful, safe, and reliable when deployed in the real world. By better evaluating summarization methods against real-world criteria, we can produce more rigorous and meaningful assessments of progress in the field.

Similarly, current benchmarks for evaluating the faithfulness of summarization models also have significant limitations that restrict their utility. These benchmarks do not accurately reflect the types of faithfulness errors made by modern state-of-the-art summarization systems. As a result, developing new evaluation metrics against these flawed benchmarks fails to produce metrics that can reliably assess faithfulness for real-world summarizers. This is a critical gap that needs to be addressed in future work. New benchmark datasets must be constructed to better represent faithfulness issues in current summarization systems. Furthermore, benchmarks must evolve as systems improve, ensuring evaluation metrics remain robust to new error modalities and distributions. Static benchmarks risk becoming outdated and ineffective. Ideally, new benchmarks should test faithfulness across axes like factual consistency, stance agreement, logical entailment, and more at both system and summary levels. As summarization expands into sensitive domains like medicine, developing rigorous faithfulness benchmarks is crucial for preventing misinformation and evaluating real-world reliability. We can produce faithfulness metrics that track progress in this critical dimension by continuously improving the evaluation benchmarks to keep pace with systems.

Our work also highlighted some challenges with human evaluation as text summarization systems become more capable. As summarizers improve in fluency and coherence, it becomes increasingly difficult for non-experts to identify faithfulness errors in the generated summaries. While current systems can produce superficially convincing summaries, they may distort key information from the source text. This underscores the need for more robust evaluation protocols beyond simple rating tasks by non-experts. Our study presented a preliminary attempt at model-assisted evaluation by providing system critiques to check summary faithfulness. However, more research is still needed to develop reliable evaluation protocols that elicit better judgments from non-expert humans with the assistance of models. For instance, future work could allow evaluators to interact with models through clarification dialogues to elicit information to make better-informed judgments. Models could highlight possible factual discrepancies or provide justifications for their summaries to assist in evaluation. Eliciting more reliable evaluation from non-expert humans via model assistance is an open area of research and is vital to improving LLM capabilities (Bowman et al. 2022).

Our LLM benchmarking work found that instruction tuning and RLHF can substantially improve performance on news summarization. However, improvements in this narrow domain do not automatically confer stronger generalized summarization capabilities, as evidenced by state-of-the-art models still struggling with faithfulness when summarizing narratives. Further research is imperative to determine the key factors for advancing broad summarization skills beyond specialized domains. One promising direction is expanding the diversity of tasks used for instruction tuning and RLHF training beyond just news articles. Incorporating datasets covering scientific papers, meetings, dialogues, narratives, and other domains could lead to more robust models. Additionally, RLHF could be leveraged to directly optimize for summary faithfulness by rewarding faithful summaries and penalizing summaries containing factual distortions. Focusing RLHF optimization on minimizing faithfulness errors may produce more trustworthy summarizers across domains. Another potential direction is exploring high-quality in-context examples for improving capabilities across tasks. Systematically assessing the impact of in-context demonstrations on unfamiliar summarization tasks could reveal their potential for bolstering generalization. Overall, progress will require analyzing model capabilities across varied datasets, combined with specialized techniques like instruction tuning, RLHF, and few-shot learning to improve generalization.

References

Akyürek, Ekin, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu (2022). *Tracing Knowledge in Language Models Back to the Training Data*.

De-Arteaga, Maria, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai (2019). "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 120–128. ISBN: 9781450361255.

Babaeianjelodar, Marzieh, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag (2020). "Quantifying Gender Bias in Different Corpora". In: *Companion Proceedings of the Web Conference 2020*. New York, NY, USA: Association for Computing Machinery, pp. 752–759. ISBN: 9781450370240.

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio (2015). "Neural machine translation by jointly learning to align and translate". In: *3rd International Conference on Learning Representations, ICLR 2015*.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. (2022). "Constitutional AI: Harmlessness from AI Feedback". In: *arXiv preprint arXiv:2212.08073*.

Bai, Yushi, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan (2022). "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback". In: *ArXiv* abs/2204.05862.

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *IEEvaluation@ACL*.

Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. (2023). "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity". In: *arXiv preprint arXiv:2302.04023*.

Barshan, Elnaz, Marc-Etienne Brunet, and Gintare Karolina Dziugaite (2020). "RelatIF: Identifying Explanatory Training Examples via Relative Influence". In: *AISTATS*.

Bartl, Marion, Malvina Nissim, and Albert Gatt (2020). "Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias". In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 1–16.

Barzilay, Regina and Kathleen R. McKeown (2005). "Sentence Fusion for Multidocument News Summarization". In: *Computational Linguistics* 31.3, pp. 297–328.

Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad (June 1999). "Information Fusion in the Context of Multi-Document Summarization". In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, pp. 550–557.

Bhandari, Manik, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig (Nov. 2020). "Re-evaluating Evaluation in Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9347–9359.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (July 2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5454–5476.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., pp. 4356–4364. ISBN: 9781510838819.

Bommasani, Rishi and Claire Cardie (Nov. 2020). "Intrinsic Evaluation of Summarization Datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher Re, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramer, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang (2021). "On the Opportunities and Risks of Foundation Models". In: *ArXiv* abs/2108.07258.

Bowman, Samuel R, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. (2022). "Measuring progress on scalable oversight for large language models". In: *arXiv preprint arXiv:2211.03540*.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *NeurIPS*.

Caglayan, Ozan, Pranava Madhyastha, and Lucia Specia (Dec. 2020). "Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 2322–2328.

Caliskan, Aylin, Joanna Bryson, and Arvind Narayanan (Apr. 2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356, pp. 183–186.

Cao, Yang, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan (May 2022). "On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 561–570.

Cao, Ziqiang, Furu Wei, Wenjie Li, and Sujian Li (2018). "Faithful to the Original: Fact Aware Neural Abstractive Summarization". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 4784–4791.

Carlini, Nicholas, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song (2019). "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks". In: *Proceedings of the 28th USENIX Conference on Security Symposium*. SEC'19. Santa Clara, CA, USA: USENIX Association, pp. 267–284. ISBN: 9781939133069.

Chen, Anthony, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh (Aug. 2021). "Evaluating Entity Disambiguation and the Role of Popularity in Retrieval-Based NLP". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 4472–4485.

Chen, Mingda, Zewei Chu, Sam Wiseman, and Kevin Gimpel (May 2022). "SummScreen: A Dataset for Abstractive Screenplay Summarization". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 8602–8615.

Chen, Sihao, Fan Zhang, Kazoo Sone, and Dan Roth (June 2021). "Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 5935–5941.

Chen, Yen-Chun and Mohit Bansal (July 2018). "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 675–686.

Chen, Yulong, Yang Liu, Liang Chen, and Yue Zhang (Aug. 2021). "DialogSum: A Real-Life Scenario Dialogue Summarization Dataset". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 5062–5074.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel (2022). "PaLM: Scaling Language Modeling with Pathways". In: *ArXiv* abs/2204.02311.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30.

Chromik, Michael, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz (2021). "I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI". In: *26th International Conference on Intelligent User Interfaces*. IUI '21. College Station, TX, USA: Association for Computing Machinery, pp. 307–317. ISBN: 9781450380171.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning (2020). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *ICLR*.

Cohan, Arman, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian (June 2018). "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 615–621.

Cohere (2022). *Introduction to Large Language Models*. https://docs.cohere.ai/docs/introduction-to-large-language-models.

Cohn, Trevor and Mirella Lapata (Aug. 2008). "Sentence Compression Beyond Word Deletion". In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, pp. 137–144.

Dai, Damai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei (May 2022). "Knowledge Neurons in Pretrained Transformers". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 8493–8502.

Dang, Hoa Trang (2005). "Overview of DUC 2005". In: *Proceedings of the document understanding conference*. Vol. 2005. Citeseer, pp. 1–12.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dhamala, Jwala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta (2021). "BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 862–872. ISBN: 9781450383097.

Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston (2019). *Wizard of Wikipedia: Knowledge-Powered Conversational agents*. arXiv: `1811.01241 [cs.CL]`.

Dong, Li, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon (2019). *Unified Language Model Pre-training for Natural Language Understanding and Generation*. arXiv: 1905.03197 [cs.CL].

Dong, Yue, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung (Oct. 2018). "BanditSum: Extractive Summarization as a Contextual Bandit". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3739–3748.

Dong, Yue, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu (Nov. 2020). "Multi-Fact Correction in Abstractive Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9320–9331.

Dreyer, Markus, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi (2021). *Analyzing the Abstractiveness-Factuality Tradeoff With Nonlinear Abstractiveness Constraints*. arXiv: 2108.02859 [cs.CL].

Du, Zhengxiao, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang (2021). "GLM: General Language Model Pretraining with Autoregressive Blank Infilling". In: *ACL*.

Durmus, Esin, He He, and Mona Diab (July 2020). "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5055–5070.

Durmus, Esin, Faisal Ladhak, and Tatsunori Hashimoto (May 2022). "Spurious Correlations in Reference-Free Evaluation of Text Generation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1443–1454.

Dušek, Ondřej, David M. Howcroft, and Verena Rieser (Aug. 2019). "Semantic Noise Matters for Neural Natural Language Generation". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 421–426.

Dušek, Ondřej, Jekaterina Novikova, and Verena Rieser (2020). "Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge". In: *Computer Speech & Language* 59, pp. 123–156.

Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter (2021). "Evaluating Groundedness in Dialogue Systems: The BEGIN Benchmark". In: *CoRR* abs/2105.00071. arXiv: 2105.00071.

Fabbri, Alexander, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong (July 2022). "QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization". In: *Proceedings of* *the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2587–2601.

Fabbri, Alexander R, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev (2020). "SummEval: Re-evaluating Summarization Evaluation". In: *arXiv preprint arXiv:2007.12626*.

Falke, Tobias, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych (July 2019). "Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2214–2220.

Fan, Angela, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel (Nov. 2020). "Generating Fact Checking Briefs". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7147–7161.

Filippova, Katja and Michael Strube (2008). "Sentence fusion via dependency graph compression". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 177–185.

Freitag, Markus, David Grangier, and Isaac Caswell (2020). "BLEU might be Guilty but References are not Innocent". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gabriel, Saadia, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao (Aug. 2021). "GO FIGURE: A Meta Evaluation of Factuality in Summarization". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 478–487.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky (2016). "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1, pp. 2096–2030.

Gao, Xiang, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan (Nov. 2020). "Dialogue Response Ranking Training with Large-Scale Human Feedback Data". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 386–395.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou (Aug. 2021). "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics". In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, pp. 96–120.

Gehrmann, Sebastian, Yuntian Deng, and Alexander Rush (Oct. 2018). "Bottom-Up Abstractive Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4098–4109.

Gehrmann, Sebastian, Hendrik Strobelt, and Alexander Rush (July 2019). "GLTR: Statistical Detection and Visualization of Generated Text". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 111–116.

Gliwa, Bogdan, Iwona Mochol, Maciej Biesek, and Aleksander Wawer (Nov. 2019). "SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization". In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, pp. 70–79.

Goldfarb-Tarrant, Seraphina, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez (Aug. 2021). "Intrinsic Bias Metrics Do Not Correlate with Application Bias". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 1926–1940.

Gonen, Hila and Yoav Goldberg (June 2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 609–614.

Gopalakrishnan, Karthik, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür (2019). "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations". In: *Proc. Interspeech 2019*, pp. 1891–1895.

Goyal, Tanya and Greg Durrett (Nov. 2020). "Evaluating Factuality in Generation with Dependency-level Entailment". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 3592–3603.

Goyal, Tanya and Greg Durrett (June 2021). "Annotating and Modeling Fine-grained Factuality in Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 1449–1462.

Goyal, Tanya, Junyi Jessy Li, and Greg Durrett (2022). "News Summarization and Evaluation in the Era of GPT-3". In: *ArXiv* abs/2209.12356.

Goyal, Tanya, Junyi Jessy Li, and Greg Durrett (2022). "Snac: Coherence error detection for narrative summarization". In: *arXiv preprint arXiv:2205.09641*.

Graff, David, Junbo Kong, Ke Chen, and Kazuaki Maeda (2003). "English gigaword". In: *Linguistic Data Consortium, Philadelphia* 4.1, p. 34.

Grusky, Max, Mor Naaman, and Yoav Artzi (June 2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719.

Gunel, Beliz, Chenguang Zhu, Michael Zeng, and Xuedong Huang (2020). "Mind The Facts: Knowledge-Boosted Coherent Abstractive Text Summarization". In: *CoRR* abs/2006.15435. arXiv: `2006.15435`.

Guo, Han, Ramakanth Pasunuru, and Mohit Bansal (July 2018). "Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 687–697.

Guo, Wei and Aylin Caliskan (2021). "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, pp. 122–133. ISBN: 9781450384735.

Han, Wenjuan, Bo Pang, and Ying Nian Wu (Aug. 2021). "Robust Transfer Learning with Pretrained Language Models through Adapters". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 854–861.

Han, Xiaochuang, Byron C. Wallace, and Yulia Tsvetkov (July 2020). “Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5553–5563.

Hara, Satoshi, Atsushi Nitanda, and Takanori Maehara (2019). “Data Cleansing for Models Trained with SGD”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Hermann, Karl Moritz, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). “Teaching Machines to Read and Comprehend”. In: *NIPS*, pp. 1693–1701.

Honnavalli, Samhita, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang (2022). *Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation*.

Honnibal, Matthew and Ines Montani (2017). “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. To appear.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). *Parameter-Efficient Transfer Learning for NLP*. arXiv: `1902.00751 [cs.LG]`.

Hsu, Wan-Ting, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun (July 2018). “A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 132–141.

Ilyas, Andrew, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry (2022). “Datamodels: Predicting predictions from training data”. In: *arXiv preprint arXiv:2202.00622*.

Jiang, Yichen and Mohit Bansal (Oct. 2018). “Closed-Book Training to Improve Summarization Encoder Memory”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4067–4077.

Jin, Xisen, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren (June 2021). “On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 3770–3783.

Jing, Hongyan (2000). “Sentence Reduction for Automatic Text Summarization”. In: *Applied Natural Language Processing Conference*.

Jing, Hongyan and Kathleen McKeown (2000). “Cut and Paste Based Text Summarization”. In: *Applied Natural Language Processing Conference*.

Jing, Hongyan and Kathleen R McKeown (1999). “The decomposition of human-written summary sentences”. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 129–136.

Jones, Karen Sparck (1998). “Automatic summarising: factors and directions”. In: *CoRR* cmp-lg/9805011.

Kang, Daniel and Tatsunori B. Hashimoto (July 2020). “Improved Natural Language Generation via Loss Truncation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 718–731.

Kedzie, Chris and Kathleen Mckeown (2019). “A Good Sample is Hard to Find: Noise Injection Sampling and Self-Training for Neural Language Generation Models”. In: *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 584–593.

Kedzie, Chris, Kathleen McKeown, and Hal Daumé III (Oct. 2018). “Content Selection in Deep Learning Models of Summarization”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1818–1828.

Keskar, Nitish Shirish, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher (2019). “Ctrl: A conditional transformer language model for controllable generation”. In: *arXiv preprint arXiv:1909.05858*.

Kirichenko, Polina, Pavel Izmailov, and Andrew Gordon Wilson (2022). “Last layer re-training is sufficient for robustness to spurious correlations”. In: *arXiv preprint arXiv:2204.02937*.

Knight, Kevin and Daniel Marcu (2002). “Summarization beyond sentence extraction: A probabilistic approach to sentence compression”. In: *Artificial Intelligence* 139.1, pp. 91–107.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes (2021). “To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation”. In: *CoRR* abs/2107.10821. arXiv: `2107.10821`.

Koh, Pang Wei and Percy Liang (Aug. 2017). “Understanding Black-box Predictions via Influence Functions”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1885–1894.

Koupaee, Mahnaz and William Yang Wang (2018). “Wikihow: A large scale text summarization dataset”. In: *arXiv preprint arXiv:1810.09305*.

Kryscinski, Wojciech, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher (Nov. 2019). “Neural Text Summarization: A Critical Evaluation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 540–551.

Kryscinski, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher (Nov. 2020). “Evaluating the Factual Consistency of Abstractive Text Summarization”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9332–9346.

Kryściński, Wojciech, Romain Paulus, Caiming Xiong, and Richard Socher (Oct. 2018). “Improving Abstraction in Text Summarization”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1808–1817.

Kryściński, Wojciech, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev (2021). “Booksum: A collection of datasets for long-form narrative summarization”. In: *arXiv preprint arXiv:2105.08209*.

Kurita, Keita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov (Aug. 2019). “Measuring Bias in Contextualized Word Representations”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, pp. 166–172.

Laban, Philippe, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst (2022). “SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization”. In: *Transactions of the Association for Computational Linguistics* 10, pp. 163–177.

Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown (Nov. 2020). “WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4034–4048.

Ladhak, Faisal, Esin Durmus, and Tatsunori Hashimoto (July 2023). “Contrastive Error Attribution for Finetuned Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 11482–11498.

Ladhak, Faisal, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown (May 2022). “Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, pp. 1410–1421.

Ladhak, Faisal, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto (May 2023). "When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3206–3219.

Ladhak, Faisal, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown (July 2020). "Exploring Content Selection in Summarization of Novel Chapters". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5043–5054.

Lai, Vivian, Chacha Chen, Qingzi Vera Liao, Alison Smith-Renner, and Chenhao Tan (2021). "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies". In: *ArXiv* abs/2112.11471.

Lakkaraju, Himabindu, Stephen H. Bach, and Jure Leskovec (2016). "Interpretable Decision Sets: A Joint Framework for Description and Prediction". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, pp. 159–174.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (Sept. 2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (June 2018). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 687–692.

Lehnert, Wendy G (1981). "Plot units and narrative summarization". In: *Cognitive science* 5.4, pp. 293–331.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880.

Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu (Nov. 2017). "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 986–995.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81.

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau (Nov. 2016). "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2122–2132.

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740.

Liu, Yixin, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq R. Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir R. Radev (2022). "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation". In: *ArXiv* abs/2212.07981.

Liu, Yixin, Pengfei Liu, Dragomir R. Radev, and Graham Neubig (2022). "BRIO: Bringing Order to Abstractive Summarization". In: *Annual Meeting of the Association for Computational Linguistics*.

Longpre, Shayne, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh (Nov. 2021). "Entity-Based Knowledge Conflicts in Question Answering". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7052–7063.

Lux, Klaus-Michael, Maya Sappelli, and Martha Larson (Nov. 2020). "Truth or Error? Towards systematic analysis of factual errors in abstractive summaries". In: *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Online: Association for Computational Linguistics, pp. 1–10.

Mani, Inderjeet and Eric Bloedorn (1999). "Summarizing similarities and differences among related documents". In: *Information Retrieval* 1.1-2, pp. 35–67.

Marcu, Daniel (1997). "From discourse structures to text summaries". In: *Intelligent Scalable Text Summarization*.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn (July 2020). “Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4984–4997.

Matsumaru, Kazuki, Sho Takase, and Naoaki Okazaki (July 2020). “Improving Truthfulness of Headline Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1335–1346.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (July 2020). “On Faithfulness and Factuality in Abstractive Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1906–1919.

McDonald, Ryan (2006). “Discriminative sentence compression with soft syntactic evidence”. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 297–304.

Mihalcea, Rada and Hakan Ceylan (2007). “Explorations in automatic book summarization”. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 380–389.

Mitchell, Eric, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning (2022). “Fast Model Editing at Scale”. In: *International Conference on Learning Representations*.

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang (Aug. 2016). “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond”. In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, pp. 280–290.

Nan, Feng, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang (Apr. 2021). “Entity-level Factual Consistency of Abstractive Text Summarization”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 2727–2733.

Nan, Feng, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang (Aug. 2021). “Improving Factual Consistency of Abstractive Summarization via Question Answering”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6881–6894.

Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman (Nov. 2020). “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models”. In:

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, pp. 1953–1967.

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018). "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium.

Nenkova, Ani, Sameer Maskey, and Yang Liu (June 2011). "Automatic Summarization". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Portland, Oregon: Association for Computational Linguistics, p. 3.

Novikova, Jekaterina, Ondřej Dušek, and Verena Rieser (Aug. 2017). "The E2E Dataset: New Challenges For End-to-End Generation". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Saarbrücken, Germany: Association for Computational Linguistics, pp. 201–206.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe (2022). "Training language models to follow instructions with human feedback". In: *ArXiv* abs/2203.02155.

Pagnoni, Artidoro, Vidhisha Balachandran, and Yulia Tsvetkov (June 2021). "Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4812–4829.

Papalampidi, Pinelopi, Frank Keller, Lea Frermann, and Mirella Lapata (2020). "Screenplay Summarization Using Latent Narrative Structure". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1920–1933.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.

Park, Ji Ho, Jamin Shin, and Pascale Fung (Oct. 2018). "Reducing Gender Bias in Abusive Language Detection". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2799–2804.

Pasunuru, Ramakanth and Mohit Bansal (June 2018). "Multi-Reward Reinforced Summarization with Saliency and Entailment". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* *Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 646–653.

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller (Nov. 2019). “Language Models as Knowledge Bases?” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2463–2473.

Pfeiffer, Jonas, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych (2020). “AdapterHub: A Framework for Adapting Transformers”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*. Online: Association for Computational Linguistics, pp. 46–54.

Phy, Vitou, Yang Zhao, and Akiko Aizawa (Dec. 2020). “Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4164–4178.

Prates, Marcelo O. R., Pedro H. C. Avelar, and Luis Lamb (2018). *Assessing Gender Bias in Machine Translation – A Case Study with Google Translate*.

Pruthi, Garima, Frederick Liu, Satyen Kale, and Mukund Sundararajan (2020). “Estimating Training Data Influence by Tracing Gradient Descent”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin.

Radev, Dragomir R., Eduard Hovy, and Kathleen McKeown (2002). “Introduction to the Special Issue on Summarization”. In: *Computational Linguistics* 28.4, pp. 399–408.

Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska (2000). “Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies”. In: *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). “Language Models are Unsupervised Multitask Learners”. In.

Rae, Jack W, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. (2021). “Scaling language models: Methods, analysis & insights from training gopher”. In: *arXiv preprint arXiv:2112.11446*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2019). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *CoRR* abs/1910.10683. arXiv: `1910.10683`.

Rajani, Nazneen, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong (2020). "Explaining and Improving Model Behavior with k Nearest Neighbor Representations". In: *ArXiv* abs/2010.09030.

Al-Rfou, Rami, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena (2015). "Polyglot-NER: Massive multilingual named entity recognition". In: *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, pp. 586–594.

Robertson, Stephen E., Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford (1994). "Okapi at TREC-3". In: *TREC*.

Rush, Alexander M., Sumit Chopra, and Jason Weston (2015). "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Salton, Gerard, Amit Singhal, Mandar Mitra, and Chris Buckley (1997). "Automatic text structuring and summarization". In: *Information Processing & Management* 33.2. Methods and Tools for the Automatic Construction of Hypertext, pp. 193–207.

Sandhaus, Evan (2008). "The new york times annotated corpus". In: *Linguistic Data Consortium, Philadelphia* 6.12, e26752.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Rose Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush (2021). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *ArXiv* abs/2110.08207.

Saunders, William, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike (2022). "Self-critiquing models for assisting human evaluators". In: *arXiv preprint arXiv:2206.05802*.

Schioppa, Andrea, Polina Zablotskaia, David Vilar Torres, and Artem Sokolov (2022). "Scaling Up Influence Functions". In: *AAAI-22*.

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 2017). "Get To The Point: Summarization with Pointer-Generator Networks". In: *Proceedings of the 55th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083.

Sellam, Thibault, Dipanjan Das, and Ankur P. Parikh (2020). "BLEURT: Learning Robust Metrics for Text Generation". In: *Annual Meeting of the Association for Computational Linguistics*.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Sinha, Koustuv, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau (July 2020). "Learning an Unreferenced Metric for Online Dialogue Evaluation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2430–2441.

Stiennon, Nisan, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano (2020). "Learning to summarize from human feedback". In: *ArXiv* abs/2009.01325.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang (July 2019). "Mitigating Gender Bias in Natural Language Processing: Literature Review". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1630–1640.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, pp. 3319–3328.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick (July 2019). "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4593–4601.

Thadani, Kapil and Kathleen McKeown (2013). "Supervised Sentence Fusion with Single-Stage Inference". In: *Proceedings of IJCNLP*. Nagoya, Japan.

Torralba, Antonio and Alexei A. Efros (2011). "Unbiased look at dataset bias". In: *CVPR 2011*, pp. 1521–1528.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way (Oct. 2018). "Getting Gender Right in Neural Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in*

Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 3003–3008.

Vasilyev, Oleg, Vedant Dharnidharka, and John Bohannon (Nov. 2020). “Fill in the BLANC: Human-free quality estimation of document summaries”. In: *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Online: Association for Computational Linguistics, pp. 11–20.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Wang, Alex, Kyunghyun Cho, and Mike Lewis (July 2020). “Asking and Answering Questions to Evaluate the Factual Consistency of Summaries”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5008–5020.

Wang, Yizhong, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi (2022). “Benchmarking Generalization via In-Context Instructions on 1, 600+ Language Tasks”. In: *ArXiv* abs/2204.07705.

West, Peter, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi (2021). “Symbolic Knowledge Distillation: from General Language Models to Commonsense Models”. In: *ArXiv* abs/2110.07178.

Whiting, Mark E., Grant Hugh, and Michael S. Bernstein (2019). “Fair Work: Crowd Work Minimum Wage with One Line of Code”. In: *AAAI Conference on Human Computation & Crowdsourcing*.

Wilber, Matt, William Timkey, and Marten van Schijndel (Aug. 2021). “To Point or Not to Point: Understanding How Abstractive Summarizers Paraphrase Text”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 3362–3376.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (Oct. 2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics, pp. 38–45.

Wu, Jeff, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano (2021). "Recursively summarizing books with human feedback". In: *arXiv preprint arXiv:2109.10862*.

Wu, Yuxiang and Baotian Hu (2018). "Learning to Extract Coherent Summary via Deep Reinforcement Learning". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 5602–5609.

Xu, Jiacheng, Shrey Desai, and Greg Durrett (Nov. 2020). "Understanding Neural Abstractive Summarization Models via Uncertainty". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6275–6281.

Yasunaga, Michihiro, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev (2019). "Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 7386–7393.

Yeh, Chih-Kuan, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar (2018). "Representer point selection for explaining deep neural networks". In: *Advances in neural information processing systems* 31.

Yeh, Yi-Ting, Maxine Eskenazi, and Shikib Mehri (Nov. 2021). "A Comprehensive Assessment of Dialog Evaluation Metrics". In: *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*. Online: Association for Computational Linguistics, pp. 15–33.

Yuan, Ann, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann (2021). "SynthBio: A Case Study in Human-AI Collaborative Curation of Text Datasets". In: *ArXiv* abs/2111.06467.

Yuan, Weizhe, Graham Neubig, and Pengfei Liu (2021). "BARTScore: Evaluating Generated Text as Text Generation". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 27263–27277.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu (2020). "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *ArXiv* abs/1912.08777.

Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston (July 2018). "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2204–2213.

Zhang, Shiyue, David Wan, and Mohit Bansal (2022). "Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization". In: *arXiv preprint arXiv:2209.03549*.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022). "OPT: Open Pre-trained Transformer Language Models". In: *ArXiv* abs/2205.01068.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi (2019). "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675*.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto (2023). "Benchmarking large language models for news summarization". In: *arXiv preprint arXiv:2301.13848*.

Zhao, Chao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi (2022). "NarraSum: A Large-Scale Dataset for Abstractive Narrative Summarization". In: *arXiv preprint arXiv:2212.01476*.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang (June 2019). "Gender Bias in Contextualized Word Embeddings". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 629–634.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (Sept. 2017). "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2979–2989.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (June 2018). "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 15–20.

Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang (Oct. 2018). "Learning Gender-Neutral Word Embeddings". In: *Proceedings of the 2018 Conference on Empirical Methods in*

Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 4847–4853.

Zhao, Zheng, Shay B. Cohen, and Bonnie Webber (Nov. 2020). “Reducing Quantity Hallucinations in Abstractive Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2237–2249.

Zhong, Ming, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang (July 2019). “Searching for Effective Neural Extractive Summarization: What Works and What’s Next”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1049–1058.

Zhou, Chunting, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad (Aug. 2021). “Detecting Hallucinated Content in Conditional Neural Sequence Generation”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 1393–1404.

Zhou, Qingyu, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao (July 2018). “Neural Document Summarization by Jointly Learning to Score and Select Sentences”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 654–663.

Zhu, Chenguang, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang (June 2021). “Enhancing Factual Consistency of Abstractive Summarization”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 718–733.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving (2019). “Fine-Tuning Language Models from Human Preferences”. In: *arXiv preprint arXiv:1909.08593*.