Aalto University
School of Science
Master's Programme in ICT innovation

Ioana Frincu

In Search of the Perfect Prompt

A User Evaluation of Soft and Hard Prompt Techniques for Conversational Abstract Generation

Master's Thesis
Espoo, September, 2023

Supervisors: Professor Lauri Juvela, Aalto University
Advisor: PhD. Ronin Wu
MSc. Victor Botev

Aalto University
School of Science
Master's Programme in ICT innovation

ABSTRACT OF
MASTER'S THESIS

Author:	Ioana Frincu		
Title			
In Search of the Perfect Prompt			
Date:	September, 2023	**Pages:**	82
Major:	HCID	**Code:**	SCI3042
Supervisors:	Professor Lauri Juvela, Aalto ELEC		
Advisor:	PhD. Ronin Wu MSc. Victor Botev		

The study investigates the efficacy of soft and hard prompt strategies in the scientific domain, namely in the tasks of conversational abstract generation. The proposed approach incorporates two distinct methods, prompt engineering and prompt tuning, within a Conversational Recommender System (CRS). The primary objective of this system is to aid users in generating abstracts for their research. The present study employs an evaluation approach that integrates user research with objective performance criteria. This study examines the strengths and disadvantages associated with both categories of prompts, commencing with an analysis of existing literature on CRS and prompting studies, and subsequently conducting original research tests. This study makes three primary contributions. Initially, a compilation of prerequisites and hypothetical situations is formed by an examination of the issue at hand. This wishlist presents a range of potential technological, user, and functional views that have the potential to contribute to future studies in this area. Furthermore, the examination of user studies is an integral element of our evaluation methodology. During this process, we analyze many factors pertaining to the 6 participants, including their cognitive load, response time, and overall happiness while applying challenging prompts within the CRS. In our investigation, we examine the behavior and needs of the target demographic, consisting of academics and researchers. Our findings suggest a tendency among this group to favor interactions that are focused on factual information and question-and-answer exchanges, as opposed to more expansive and conversational encounters. Thirdly, our study delves into the comprehensibility and relevance of the generated abstracts, utilizing well-established criteria such as Rouge and F1 scores. In our research, the anticipated effect of combining prompts with text-generation tasks is to produce scientific abstracts that are imprecise and broader in nature. However, this objective contradicts the expectations of the users. The research findings shed light on the difficulties and advantages that arise from implementing prompting techniques with a CRS. This study makes a valuable contribution by recognizing the importance of contextual comprehension and employing prompting strategies from both technical and user-centric viewpoints. One of the primary findings is that it is crucial to customize prompt tactics in accordance with user preferences and domain demands. The given findings contribute to the existing body of knowledge on conversational recommender systems and their applications in the field of natural language processing.

Keywords:	Prompting, NLP, LLM, User evalutation, HCI
Language:	English

Acknowledgements

I extend my heartfelt gratitude to several individuals and groups who helped enormously in the completion of this thesis.

First and foremost, I wish to express my sincere appreciation to my supervisor, Lauri. His unwavering guidance, deep understanding, and invaluable mentorship have not only enriched my research journey but have also imparted the essence of contributing meaningfully to the realm of science. I am especially thankful for the moments of comic relief marked by Lauri's witty jokes during our meetings, which added a delightful touch to our collaboration. Also, big thanks to Mariet Theune, my UTwente observer, for the detailed feedback and who was also one of the main reasons I chose to pursue the EIT Master.

I am profoundly indebted to my advisor, Ronin, whose exceptional support played a pivotal role in bringing this research to fruition. Alongside Victor, my second advisor, Ronin exemplified what it means to be valued within a remarkable organization and taught me "it is better to struggle together than alone". To all the members of Iris.ai, particularly the Research team, thank you for the remarkable experiences and the cherished memories. I will fondly miss our vibrant Research "ketchups".

My deepest thanks extend to my family, friends, and the EIT community, who stood by my side throughout this research journey. They witnessed the highs, lows, and the moments in between, offering unwavering support. A special tribute goes to my mother, whose steadfast belief in my abilities and encouragement to pursue my passions have been a driving force since my childhood.

This journey has been a collective endeavor enriched by the wisdom, support, and camaraderie of these exceptional individuals and groups. I am profoundly grateful for their contributions to this endeavor, and I look forward to the continued pursuit of knowledge and innovation in the years to come.

Espoo, September, 2023

Ioana Frincu

Contents

Chapter 1

Introduction

In the present era of digital advancements, the rapid expansion of accessible information has generated a critical requirement for sophisticated procedures aimed at extracting the most relevant content. This master's thesis does a thorough investigation, intended to evaluate the effectiveness of soft and hard prompt strategies in the field of generating conversational abstracts and prioritizing source links. Prompting techniques are considered crucial components in the field of natural language processing(NLP) and artificial intelligence(AI) (72). They have a significant influence on the methods used for information retrieval and summarization. The importance of doing this inquiry becomes evident when considering the increasing difficulty in producing abstracts of high quality and effectively generating source links for the scientific and research domain.

Recently, significant progress has been witnessed in NLP and AI, empowering large language models (LLM) to produce coherent and contextually relevant responses when prompted. However, there are still challenges remaining in scientific text processing. Researchers and scholars encounter difficulties efficiently accessing and summarizing vast amounts of scientific literature, impeding the topic assimilation of new knowledge and slowing the research progress. The issue stems from a lack of a practical and user-friendly conversational recommender system tailored to the scientific domain. The methodology employed in this study integrates state-of-the-art content recommendation algorithms (88) in order to conduct a comprehensive evaluation of this pivotal feature.

The potential outcomes of this study have ramifications for the fields of information retrieval, content summarization, and NLP. The objective of this study is to provide a comprehensive analysis of the distinct advantages and disadvantages associated with soft and hard prompts, thereby enhancing our comprehension of their significance in CRS that prioritize user needs (73). Our objective is to narrow the divide between human interaction and automated content generation by harmonizing prompt techniques with user preferences.

This research attempts to explore the integration of a prompt-tuned conversational recommender for scientific papers and how the users interact with it. The conversational recommender is constructed as an additional feature of the existing platform of Iris.AI, having the central role of smoothing the user adoption process by overcoming two main impediments. First, to create the query that will narrow down the topic of the relevant papers search, the user has to input a research question and a 300 to 500 word scientific abstract, which should include specific abbreviations or terminology, research scope, and methodology. The main issue arising from this user flow is the assumption that the user already has a clear definition of the research topic and is looking for related articles.

Expanding this user requirement to include users who have not yet clearly defined the research topic and are currently in an exploration phase will broaden the target audience and remove usability problems. Second, after the abstract is generated and the database of articles is generated, the conversational recommender will serve as a filtering option in answering the most relevant articles based on the user's input. Thus, the second part fulfills the tasks of informational retrieval through ranking.

Research question: "How does the implementation of prompt engineering and prompt tuning strategies in a language model conversational recommender system, specifically tailored to scientific abstract generation, impact the system's accuracy, relevance, and user satisfaction in supporting researchers' information needs and enhancing productivity?"

The primary purpose of this system is to include users in interactive conversations to provide succinct abstracts and appropriate references to relevant sources related to the academic and research literature domains. Our methodology involves the utilization of two separate prototypes. The first prototype utilizes a hard prompting technique, incorporating two basic models, namely OpenAI's LLMs such as gpt3. On the other hand, the second prototype adopts a soft prompting method, using SciBert, GPT-2, and Falcon. The models undergo minimal fine-tuning to enhance their ability to support natural and coherent user interactions. Our research aims to do a comparative analysis of these two methodologies, both qualitative and quantitative, with a specific emphasis on evaluating their performance in abstract generation and source link prioritizing. Integration of the CRS system with the text generative abilities of LLMs is the backbone architecture for this research, with the goals of evaluating both the human interaction aspect and the algorithmic performance. These findings have significance for advancing LLM based conversational recommender systems that prioritize user needs and preferences in specific context domains.

Chapter 2

Background

This literature review aims to explore and synthesize the existing body of research relevant to the language models used in conversational recommender and prompting techniques. An important aspect to consider is that this area of NLP is relatively young, with lots of research papers being published during the writing of this thesis. Thus, the literature review will mainly include published articles and research papers. By presenting a comprehensive analysis of the existing literature, the study establishes the foundation for the research questions and objectives. The literature review also highlights the gaps and limitations in the current knowledge,, highlighting the developing novel contributions of this research to the field. The literature review is structured in five subsections.

2.1 Conversational recommender systems

A Conversational Recommender System (CRS) is a system that can predict users' dynamic preferences through dialogue context to complete recommendation tasks. CRS have emerged as valuable tools for assisting users in accessing relevant information and making informed decisions. The main benefit of CRS is alleviating information overload in the users' decision-making process by simulating natural language. As an integral part of Natural Language Processing (NLP) research, conversational recommender systems have garnered significant attention due to their potential to enhance user experience and facilitate efficient information retrieval (71). E-commerce businesses, as well as tech companies, have been developing CRS for various fields, such as Amazon Alexa, Spotify, Netflix, and Starbucks Virtual Barista.

CRS can employ various NLP techniques, such as natural language understanding, dialogue management, gradient boosting decision trees, information retrieval, and neural collaborative filtering to engage users in interactive conversations and provide personalized recommendations based on their preferences, contexts, and historical interactions (25). The goal of CRS is to extract relevant information from large datasets and generate appropriate recommendations. This personalized and conversational approach enhances the user experience by tailoring offers to individual preferences and adapting to dynamic user contexts.

There has been a renewed interest in CRS and their versatile commercial applications based on the significant shift of attention towards large language models (LLM). In the recent research, CRS has been deployed using pre-trained language models (17) and evaluated on both *conversational* and *recommendation* abilities. However, newer text generative models such as ChatGPT are missing a robust evaluation on the informal recommendation tasks.

Challenges in CRS

There are two critical challenges when constructing CRS. First, as mentioned in Ricci, Rokach, and Shapira (75), the optimal exploitation of the dataset and knowledge sources of recommender systems is highly dependent on the recommendation technique. Secondly, a critique of mainstream CRS is the one-shot dialogue flow, where the user can not express any feedback on the recommendation or select alternatives, and the designs of the systems' architectures are not dynamic and integrated enough to refine the system's answers based on user responses (99; 60). Furthermore, Manzoor et al.(60) specify that the design of such systems is rarely described in scholarly papers under one unified framework, mainly due to a strict separation between the technical part being separated from the human-computer interaction module. This leads to minimal engagement with users as well as limited evaluation frameworks of recommender systems, which either exclude metric scores or user evaluations.

Papers such as Zangler and Bauer's(98) *Framework for Evaluating Recommender systems* and Zhang's (99) *System Ask-User Responds* paradigm showcase some of the attempts in offering which offer possible solution to the challenges mentioned above. However, there are still unanswered questions since CRS is a relatively young field, even though it has shown significant improvements in recent years due to fast user adoption, more extensive computational resources and new developed NLP techniques. A relevant aspect of CRS, which will be part of the core of this research, is context adaptation. Since the domain of the knowledge source affects the recommendation behavior. Defining the dialogue structure and having grounded truth for evaluation are considerations to be taken into account when developing dynamic and adaptive CRS (71). According to Pramod, "feedback from the users is critiquing and narrowing the search result".

Several conversational recommender systems have been developed for scientific literature, each with its functionalities, limitations, and user feedback. These systems employ techniques such as deep learning models, knowledge graphs, and natural language understanding algorithms to understand user queries, retrieve relevant articles, and generate informative summaries. For example, Semantic Scholar's AI-powered chatbot, known as the Allen Institute for AI (AI2) Assistant (1), assists researchers in finding relevant scientific papers by understanding natural language queries and providing personalized recommendations. Another system, SciBot, leverages deep learning models to generate abstractive summaries of scientific articles, enabling researchers to grasp the main ideas and findings quickly.

User feedback plays a crucial role in improving and refining conversational recommender systems. User studies and evaluations provide insights into these systems' usability, effectiveness, and satisfaction. By analyzing user feedback, system developers can identify areas for improvement, address limitations, and enhance the overall user experience.

However, conversational recommender systems for scientific literature also face challenges and limitations. These systems heavily rely on the availability and quality of scientific article datasets, which may vary across research domains. Furthermore, ensuring the accuracy and relevance of recommendations and generated summaries remains a challenge, as scientific articles' semantic complexity and technical nature require careful consideration.

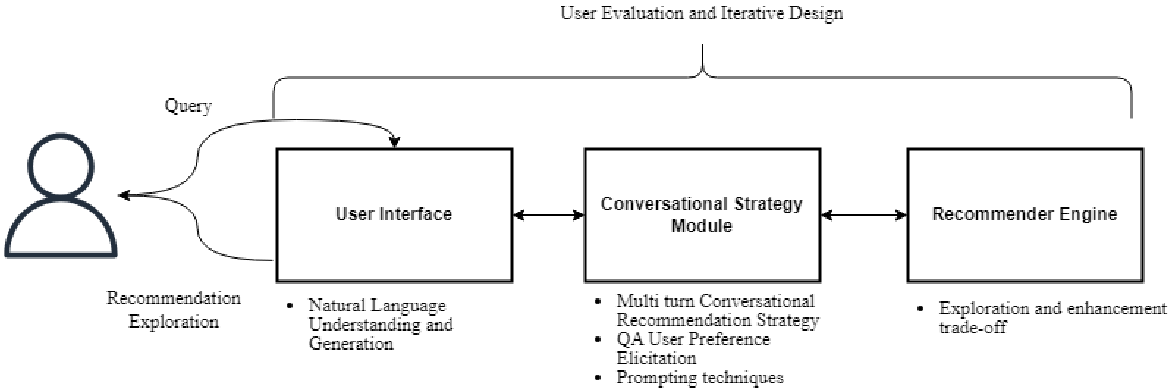

Figure 2.1: Framework of the CRS adapted from Gao's(25) framework

Figure 2.1 showcases the CRS framework adapted from Gao et al. (25) to better illustrate the techniques used in this research. The user has the option to explore diverse topics before getting the final recommendation, with the risk of not having an accurate, in-depth recommendation, or the other option, enhancing the recommendation through conversation on one topic. Gao et al. (25) mention the importance of user simulation besides the user evaluation, however, this research focuses on iterative design as a priority due to the refinement needed in implementing prompting techniques.

Evaluation of CRS

Evaluating the effectiveness of conversational recommender systems in supporting researchers' information needs is essential to assess their practical value. Several evaluation metrics and methodologies can be employed to measure the performance and user satisfaction of these systems.

Evaluation metrics commonly used in the field of conversational recommender systems include precision, recall, F1 score, and user engagement metrics (e.g., click-through rate, session duration). Additionally, qualitative assessments through user studies, surveys, and interviews can provide valuable insights into researchers' perceptions of the system's usefulness, effectiveness, and overall user experience.

To ensure the appropriateness of evaluation methodologies for scientific abstract generation, domain-specific evaluation criteria can be considered. Metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (46) and Bilingual Evaluation Understudy (BLEU) (67) are commonly used to assess the quality of generated abstracts in terms of their similarity to human-written summaries. Additionally, involving domain experts to provide subjective evaluations and comparisons can offer valuable insights into the abstracts' informativeness, coherence, and adherence to scientific standards.

In conclusion, conversational recommender systems have the potential to significantly support researchers in accessing and summarizing scientific literature. However, their effectiveness and usability need to be evaluated using appropriate metrics and methodologies. Through rigorous evaluation and user feedback analysis, researchers can identify areas for improvement and contribute to the development of more efficient and user-friendly conversational recommender systems in the scientific research domain.

2.2 Prompting techniques in NLP

In the field of Natural Language Processing (NLP), prompting strategies have developed as influential tools that direct the responses of language models. These prompts, which encompass a variety of inquiries and discussions, play a crucial role in influencing exemplary conduct. The importance of these models extends to a range of NLP applications, encompassing conversational agents, summarization techniques, and recommendation systems.

This section offers a comprehensive examination of the underlying principles and practical implementations of prompting approaches. The introductory section establishes the foundation for our thesis, with a specific emphasis on the assessment conducted by users about soft and hard prompts in open-source LLM. Our study focuses on examining the influence of these factors on the generation of conversation-based abstracts and the retrieval of pertinent source links, so making a valuable contribution to the advancing field of NLP.

Prompt engineering is a recent technique used to guide language models in generating desired outputs by providing specific instructions or cues in the form of prompts. Prompt engineering goes hand in hand with the user experience. The user and the machine benefit from this input structuring technique. Converting tasks into a language model format builds intuitiveness without further fine-tuning of the model side while the user is obtaining more human-like, authentic and meticulous responses. The effectiveness of prompt engineering lies in its ability to guide the model's decision-making towards generating desired outputs

Prompt engineering (37) aims to bias the model towards desired responses and encourage specific behaviors and actions. The output of prompted models score higher than unprompted models due to the generated answers being contextually narrowed down, more precise and coherent as compared in the article by Zhang et al. (100). The prompts can be designed to influence the tone, style or genre of the generated text, generate context-aware responses, present formulation alternative, control the level of detail, for example by giving the model an expert role provide constraints for the context, support application or domains such as code generation, reduce biases and harmful content and so on. Knowledge probing is one of the popular tasks evaluated in prompt research in order to prove that the model is learning rather than learning recall (25; 10; 72). A relevant example is Petroni et al. (70) research which demonstrated that it is possible to retrieve world facts from a pre-trained language model by formulating them as cloze-style prompts (fill-in-the-blanks format) and interpret the model's prediction accuracy as the minimum amount of factual information it contains. Jiang et al. (35) highlight that models are highly sensitivity to improperly constructed contexts which can cause artificially low perfomance. Thus, prompting techniques are useful in defining the contextual boundaries for language models.

There are two distinct prompt categories, manual and auto generated prompts.

Hard prompts (also called manual prompts) are designed by humans in probing the PLM (70). A good examples of prompts databases is *AwesomeChatGpt* github repo[1] which includes a variety of manual prompts examples that users can use to condition the context within the language model's behaviour. This is an accessible technique to all types of users that is becoming useful in building a communication framework between

[1]https://www.awesomegptprompts.com/

users and language models. However, manual prompting requires a degree of expertise from the designer and they might not be the most efficient for complex downstream task (44; 49; 54). For example, the Pattern Exploring Training (PET) model performs well on given clozed questions only in supervised and semi-supervised training on a limited dataset (81). This approach is suitable for scenarios where a single-shot interaction suffices to generate the desired output (72; 11).

When it comes to the refinement of hard prompts, there are several approaches which can be combined for achieving the desired output. Instruction prompting refers to giving a task to the model to solve. Seemingly simple approach, it still requires a bit of thought such as removing personal identification items if asked to rewrite a an email used for company's, or giving strict pointers for the evaluation of a text. Role prompting is where the prompt is giving the generative model an actual role such as food critic, teacher, doctor, writer, so that the output is styled and enhanced according to the role. Role prompting can also improve the accuracy of the output(96). Few-shot prompting is a technique which requires the inclusion of examples in the prompts which the model can used to adjust the output to the desired format. It is the preferred strategy over zero-shot and one-shot techniques because it yields more accurate results(101). All these approaches can be combined in refining the hard prompt over a few iterations. Understanding the structure of a prompt is crucial in utilizing language models effectively. The key parts of a prompt include a role, an instruction or task, a question, context, and examples. Not all prompts contain all these elements, and their order can vary. Placing the instruction last ensures the model focuses on executing the task rather than extending the context. In the field of prompt engineering, these principles provide a solid foundation for crafting effective prompts(95).

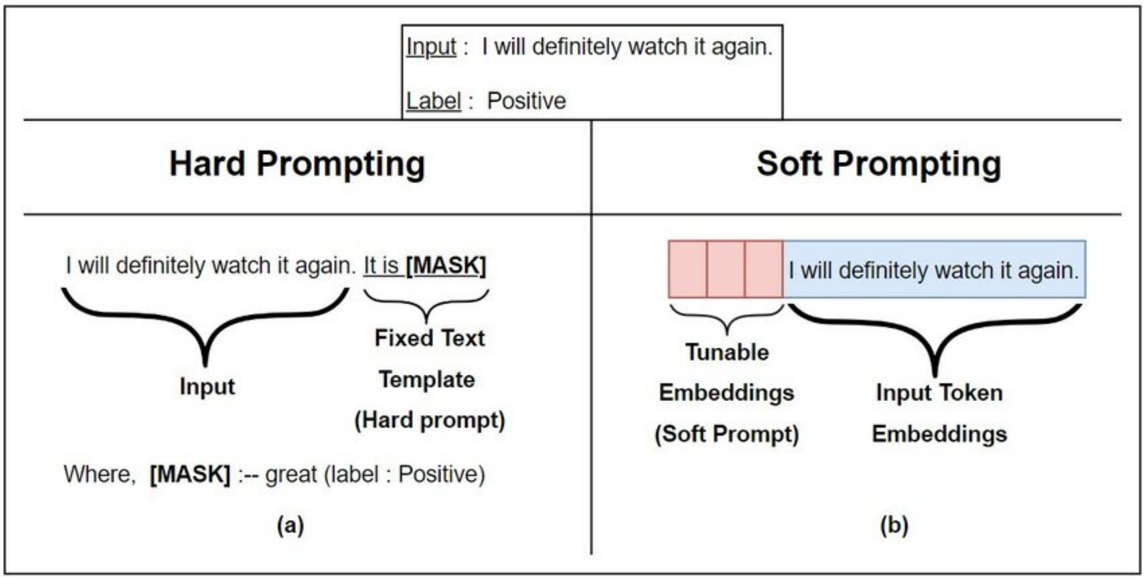

Figure 2.2: Figure from Senadeera's(82) comparison research

The later prompting approach, **soft prompting** (also called automated prompting), refers to the prompt optimally found by the algorithm through searching over the embedding (26). The automated approach can overcome the limitations of manually designed prompts by training the model to learn from a global prompt for each type of tasks such as factual probing for a specific subject-action relation, regardless of the inputs. Latest research include AutoPrompt (83), PromptGen(100) and Optiprompt (103) which automatically generate prompts based on inputs by leveraging the pre-trained generative model. For example, AutoPrompt is applied in a masked filling model such as BERT for a task by taking the task inputs in combination with trigger tokens applied to all inputs and chosen through a gradient-based search.

For a simple sentiment analysis task on a product review database as presented in Gao et al., (26), it is necessary only to design a template ("It was, ", "I think") and the expected text responses (such as positive and the negatives label). The gap between the two stages, MLM pre-training and finetuning on sentiment classification, deploying the pre-trained models on specific tasks becomes much easier, especially for the few-shot case when there are a limited number of examples for the task training(82). The process of fientuning PTM and the task specific parameters is much smoother with prompting. Scao and Rush(80) show that a prompt may be worth 100 conventional data points, suggesting that prompts can bring a giant leap in sample efficiency.

Therefore, it is essential to clearly distinguish between prompt engineering and prompt tuning. Prompt engineering refers to the refinement of manual prompts over time in order to achieve the desired output. In contrast, prompt tuning is an alternative to fine-tuning a model, where prompts are embedded in one of the model's layers of self-attention. Continuous prompts involve providing a continuous stream of input to the model during generation, allowing for a dynamic and interactive conversation. This approach enables users to refine their queries or add additional context as the conversation progresses, leading to more contextually relevant and coherent responses (85; 19).

Both hard and soft prompts offer unique advantages based on the context and user requirements. Soft prompts excel in conversational settings, enabling a back-and-forth exchange with the model for more refined results. Hard prompts are efficient for one-time, non-interactive tasks where a fixed prompt is sufficient to elicit the desired response. Research has recently explored both prompt engineering and prompt tuning techniques to enhance the performance and usability of language models in various natural language processing tasks (11; 55).

Prompting approaches

Researchers have explored various aspects of prompt engineering and prompt tuning, including prompt formats, length, auto-prompting generation and information density. An overall literature overview of the prompting paradigm in combination with pre-trained models is encapsulated in Liu, Zhang and Gulla's framework, (50) *Language Modelling Paradigm Adaption for Recommender System*. The paradigm's contributions consist of optimizing pre-training models by introducing prompting in fine tuning downstream objects and increasing efficiency due to lower number of parameters needed for prompt engineering.

Prompt engineering techniques have been widely employed in various NLP tasks, including text completion, question answering, and dialogue generation. In text completion, researchers have used prompts to influence the generated text's style, tone, or topic. For example, Keskar et al. (38) explored the use of control codes in prompts to generate text with specific attributes, such as sentiment or politeness. In question answering, prompts are designed to provide context and guide the model in generating accurate and informative answers (9).

In dialogue generation, prompts play a crucial role in maintaining coherence and engaging in meaningful conversations. Zhang et al. (92) proposed a method called

Incremental Prompting, which enables the incremental generation of dialogue responses using carefully designed prompts that capture the dialogue history and context. However, as seen in Zheng (102), conversational-based prompt are the most difficult to ground and evaluate, compared to discrete and continuous prompts.

Context management is another crucial aspect of prompt engineering, as it ensures the model maintains a consistent understanding of the conversation and generates appropriate responses. Strategies such as context concatenation, attention masking, or context gating have been proposed to manage and incorporate conversation history into the prompt (66).

For prompt refinement purposes, researchers often conduct automatic and human evaluations to assess the quality, relevance, and coherence of the generated text and refine the prompts accordingly based on feedback and evaluation metrics. An in-depth discussion on evaluation methods and metrics can be found in Chapter 6. Overall, there are different approaches and strategies for prompt design, including single-turn and multi-turn prompts, context management, and prompt refinement.

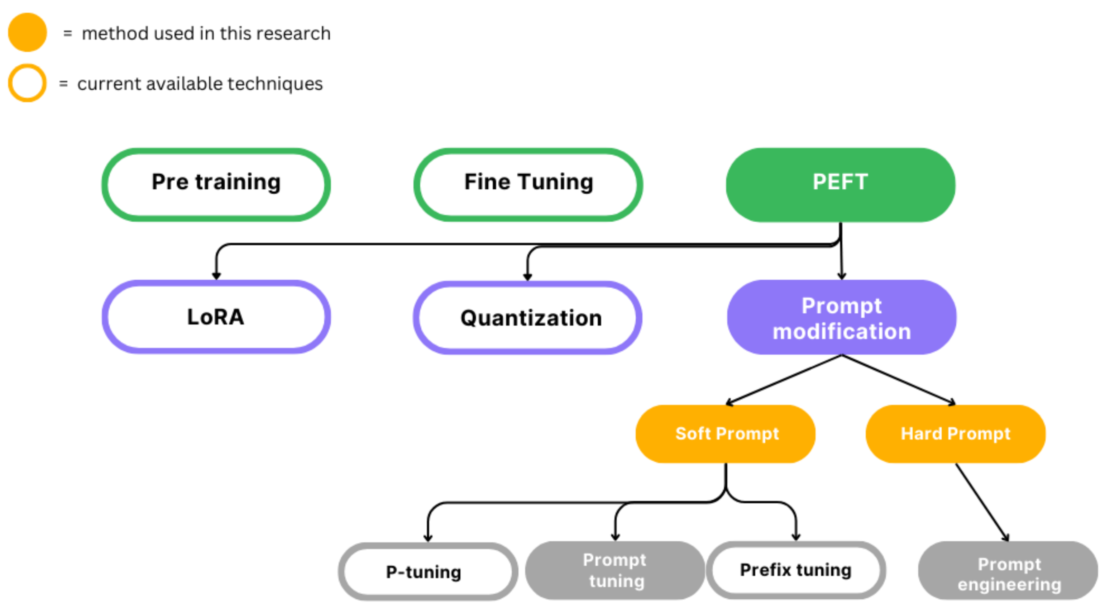

Figure 2.3: Prompting techniques

One strategy that stands out prominently among the options is Parameter-Efficient Fine Tuning (PEFT) (42). The notable characteristic of PEFT is its utilization of adapter layers, which provide an effective approach for fine-tuning pre-trained language models without requiring the retraining of the entire architecture. This methodology offers a synergistic combination of model adaptability and resource efficiency, rendering it progressively advantageous in practical scenarios. Moreover, in the realm of fine-tuning, we come across Prefix Tuning. Li and Liang (45) proposed a method called Prefix-Tuning, where explicit prompts are added at the beginning of the input text to influence the model's behavior. This approach allows fine-grained control over the generated outputs and enables the model to conform to specific requirements (15; 6).

In addition to fine-tuning, soft prompting techniques have gained prominence by introducing a dynamic and contextually aware aspect to interactions with language models. The utilization of soft prompts in dialogue systems enables the generation of responses that exhibit a higher degree of adaptability, as they are closely connected to the preceding conversation. Simultaneously, Prompt Tuning (82) has garnered acknowledgment as a meticulous method for fine-tuning, with a specific emphasis on the manipulation of prompts to optimize the fine-tuning process for targeted tasks (51; 94).

In addition to the above discussion, P-Tuning (52) is a method that involves the integration of pattern-based prompts into prompt engineering, allowing for the dynamic adjustment of prompts based on the input context. The capacity to adapt enables P-Tuning to handle a wide range of conversational circumstances effectively (35; 89).

To optimize the efficiency and administration of the models, we utilized a technique known as Parameter-Efficient Fine Tuning (PEFT), which falls under the umbrella of prompt tuning. According to Lester et al. (42), using PEFT not only resulted in reduced computational requirements but also enabled a more expedited adjustment to particular conversational contexts. PEFT was implemented via Learnable Optimizer-Ready Adapters (LORA) (33). The presence of adapter weights in LORA, a versatile component, facilitated the seamless integration of these weights, hence augmenting the flexibility and adaptability of the models in accommodating diverse conversational circumstances.

In the academic context, adapters emerged as an early parameter-efficient fine-tuning technique. This method enhances the transformer architecture by adding adapter layers, which are fine-tuned explicitly without retraining the entire model. Research findings demonstrate that this approach achieves performance levels comparable to comprehensive fine-tuning but with significantly reduced computational demands and training time. The adapter layer design reduces input dimensionality, applies non-linear activation, and scales the output to maintain compatibility with subsequent layers. The efficiency of this approach is highlighted by a remarkable 0.4% improvement on the GLUE benchmark, achieved with a mere 3.6% increase in parameters (32).

2.3 Language models for scientific abstract generation, summarization and information retrieval

Transformer-based language models have shown impressive capabilities in various natural language processing tasks. These models, which rely on extensive pre-training using diverse text datasets, excel in capturing contextual information and understanding complex linguistic patterns. In scientific text processing, these language models have become a popular choice for tasks such as abstract generation, summarization, and information retrieval. Their ability to generate coherent and informative abstracts from scientific articles has been widely recognized, making them valuable tools for researchers and practitioners in the scientific community. One such pre-trained model is SciBERT (5). The paper provides empirical evidence for the effectiveness of transformer-based language models in handling scientific text with an F1 score of 90% on the Name Entity Recognition (NER) task. It highlights the significance of domain-specific pre-training and showcases the improvements achieved by specialized language models for scientific abstract generation and related tasks.

There are two similar tasks we need to distinguish between, generating a scientific abstract and summarizing a text. Before we dive deeper into them, we should clarify that while both tasks involve creating concise representations of longer texts, summarization focuses on extracting and condensing existing content from a source text, while abstract generation may involve creating new content to summarize the core points of a research paper or article.

First, *Abstract generation* plays a critical role in scientific literature as it provides a concise summary of a research paper's main contributions. Language models can be fine-tuned on scientific literature datasets to generate informative and coherent abstracts that capture the essence of the underlying research. While this task can involve highlighting the paper's content and generating a concise summary that accurately represents the key findings, methodology, and significance of the study, it can also generate new content based on little information such as one sentence. While we can not expect generated abstracts to have in depth interpretation of results, it can use the knowledge of pre-trained models (PTM) to create a possible scientific abstract for any topic following the key structural elements of most of the research papers. One study by Ermakova et al.(22) proposes an evaluation metric for the abstract generation called GEM. The results show that recent publications exhibit a trend of more comprehensive abstracts, and it appears that there is no discernible correlation between the GEM score and the citation rate of the papers.

Second, *Summarization*, a related task, involves generating condensed versions of articles while preserving the crucial information. Language models can be employed as a solution to produce abstract summaries that capture the main points and salient details of a scientific paper. However, some of the main challenges of automatic summarization are semantic analysis and discourse analysis, where both qualitative and quantitative methods need to be used in order to accurately assess the abstract's coherence and factuality. Thus summarization tasks can not be used to generate novel content on minimal information. Current evaluation protocols for summarization tasks do not account enough for factuality (40).

One relevant tasks connected to both text generation and summarization is *information retrieval* (IR) in scientific text processing which focuses on retrieving relevant research papers given a query or specific requirements. Language models can be utilized to index and search scientific literature repositories, enabling efficient retrieval of relevant papers based on keywords, topics, or contextual similarity. Feldman (23) specifies that NLP techniques can be added in any of the four steps of IR (document processing, query processing, query matching and ranking and sorting).

Overview of LLM

Transformer-based models, such as Generative Pre-trained Transformer (GPT) (24) and Bidirectional Encoder Representations from Transformers (BERT) (18), have demonstrated outstanding performance in various NLP tasks, including scientific text processing. These models leverage self-attention mechanisms to capture contextual dependencies and generate high-quality text.

GPT, in particular, has been widely adopted for abstractive summarization and abstract generation tasks. Researchers have fine-tuned GPT on scientific literature datasets to generate coherent and informative abstracts. By conditioning the model on the input paper, GPT can generate abstracts that capture the essential information and maintain the coherence and fluency of scientific writing.

BERT (18), known for its bidirectional contextual representations, has been applied to scientific text processing tasks, including abstract generation. BERT-based models employ a combination of encoder-decoder architectures and self-attention mechanisms to generate abstractive summaries. By leveraging BERT's pre-trained representations, these models can capture the contextual nuances of language and generate abstracts that align with the original paper's content. There are several BERT variations adapted to various domains and and trained for one or more tasks, providing unique characteristics and performances for specific use cases. These adaptations may involve incorporating domain-specific knowledge, leveraging additional features or metadata, or applying reinforcement learning techniques to improve the quality and transfer learning on different domains (31; 64; 14).

Overall, transformer-based models, such as GPT and BERT, have shown great promise in scientific text processing tasks, including abstract generation. Further research and advancements in these models and domain-specific adaptations can contribute to more accurate and informative outputs.

Challenges in scientific abstract generation

The field of language models for scientific abstract generation has witnessed extensive research efforts, uncovering both advancements and challenges. Researchers have explored various techniques, ranging from extractive to abstractive methods, to generate informative abstracts from scientific papers. Extractive methods involve selecting sentences or passages directly from the original paper as the abstract. These methods employ sentence salience scoring techniques, such as PageRank or sentence embeddings, to identify the most important sentences that summarize the content. While extractive methods ensure factual accuracy, they often result in redundant or fragmented summaries that lack overall coherence, as observed in LexRank (21). In contrast, abstractive methods leverage language models to generate novel abstracts by paraphrasing and rephrasing the original paper's content. These methods offer more flexibility in summarization but face challenges in generating coherent and contextually appropriate abstracts. Ensuring the fidelity and cohesion of the generated abstracts remains a significant challenge in abstractive approaches (77)

Additionally, the scientific domain's complexity and technical language pose challenges for language models. Scientific texts often contain domain-specific jargon, acronyms, and specialized terminology, which can hinder accurate representation and understanding by the models. The challenge lies in effectively capturing the nuanced language patterns specific to scientific literature (58; 57).

Another challenge in the field is the limited availability of high-quality annotated datasets for training and evaluating language models in scientific abstract generation. Annotated scientific abstract datasets are crucial for model development and benchmarking, but their scarcity limits research progress and hinders comparative evaluations of different techniques (45).

Addressing these challenges requires further research and innovation in developing models that better understand and generate coherent and informative abstracts from scientific literature.

2.4 Related work on prompt engineering in language models for conversational recommender systems

Prompt engineering has been extensively studied in the context of conversational recommender systems, with a growing body of research. This section provides a comprehensive review of the existing literature on prompt engineering for conversational recommender systems, emphasizing its relevance and effectiveness in the scientific research domain. Additionally, a comparison of different prompt engineering approaches employed in related studies highlights their respective strengths, weaknesses, and potential for adaptation to the current research. A comparative analysis was done by Liu. et al (50) of the prompt engineering approaches employed in related studies provides valuable insights into their strengths, weaknesses, and potential applicability to the current research on prompt engineering in a language model. The prompting paradigm is flexible and adaptable and combines well with pre-trained models. Including prompting techniques in the pre-trained model itself can directly make it able to predict next items, generate recommendation explanations, make conversations or even output subtasks related to recommendation targets such as explanations. Liu et al. separate the Language Model Recommendation systems in two main parts, *Training Strategies* and *Learning Objectives*. Both of them will be addressed in the 3 section (Methodology) with regards to this research. Also, worth mentioning that each approach of the prompting paradigm has its own merits and limitations, and their applicability to the current research should be carefully considered. Based on the unique characteristics of scientific articles and the requirements of abstract synthesis, a hybrid approach combining template-based prompts with context augmentation techniques and prompt tuning may offer a balanced solution, leveraging the strengths of each method while mitigating their weaknesses.

A review of the literature reveals several notable studies that have investigated the role of prompt engineering in conversational recommender systems taking into account the interaction with the user. Manzoor and Jannach (60) research suggests that *retrieval-based approaches*, a topic that has received limited attention in the context of Conversational Recommender Systems (CRS) literature, hold potential as promising alternatives to generation-based approaches. The retrieval process typically involves matching the user's query or input against a database of predefined responses using similarity metrics like cosine similarity, word embeddings, or other semantic matching algorithms. Combining these two approaches in a hybrid system could offer a way to address the limitations of each method, leading to more reliable systems capable of effectively responding to user utterances. A challenge highlighted in the study is that chatbot would rather respond with a movie recommendation even to specific queries even if it is incorrect. Iovine et al. (34) study highlights that when it comes to natural language in CRS, it is essential to identify the most stressful activities for the user and design the CRS's goal in such a way it addresses it in an accurate manner, even at the cost of the system inquiring more about the topic. Furthermore, Meena chatbot (55) also struggles with the downside of a retrieval-based approach when it comes to out-of-domain queries. Its human likeness score is high in conversational settings, however the knowledge-based and perplexity are limited thus making it difficult to adapt to specific and more complex domains. Open-domain chatbots are still a challenge from both technical and interaction perspectives. There are studies that focus solely on user interaction aspect of a chatbot for a specific domain which can offer insights into better evaluation protocols for user experience and satisfaction (63; 76; 44), while others bring into the conversation the integrity of data the model is trained on and possible bias issues (76; 20) and some focus on frameworks which integrate better the search and conversational aspect of CRS(99). One solution to the knowledge-flexibility tradeoff of CRS is proposed by Liu et al. (48) which uses generated knowledge prompts to inject common sense reasoning into specific tasks instead of fine-tuning the entire model.

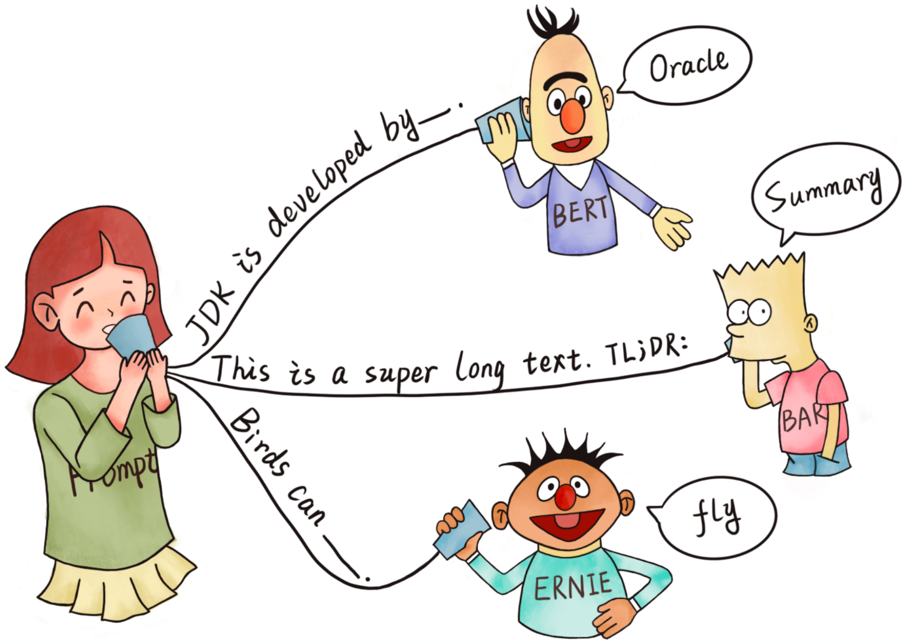

Figure 2.4: Image created by Liu et al.(49)

Prompting techniques can help in shortening the distance between the human-like aspect and the factuality and coherence of the system's responses. Recent work by Wang et al. (91) explored prompt engineering techniques for conversational recommender systems in the biomedical domain. They investigated the impact of different prompt styles, such as declarative, interrogative, and imperative prompts, on the quality of recommendations for biomedical articles. Their results indicated that prompt engineering played a vital role in tailoring recommendations to users' needs and preferences in the scientific research domain. Furthermore, Chain-of-Thought prompting (93), prefix tuning (45) and few shot prompt tuning for soft prompts (43; 51; 102) are all techniques which show the same performance with fine-tuning. The main advantage of prompt tuning is the benefit in robustness to domain transfer (15) and scaling down of the parameters (54).

Gupta et al. (30) offer an overview of the abstractive summarization current state on scientific papers. Abstractive summarization is "a technique in which the summary is generated by generating novel sentences by either rephrasing or using the new words instead of simply extracting the important sentences". Over the years, different abstractive summarization techniques have been explored in specific branches of scientific domains (13; 21). Liu et al. (51) succeed in optimizing the summarization task by conducting prompt pre-training with self-supervised data, followed by fine-tuning the model with few-shot examples of prompting. Another model, SPeC (16) uses both soft and discrete prompts to tackle the issue of performance variability in prompt tuning with regards to summarization of clinical text. As noted in several studies, essential topics when discussing the NLP techniques such as summarizations, especially when processing scientific text, are the informational retrieval and word embedding steps. Several studies are optimizing the search process within documents as well as the similarity between words so that semantic meaning is not compromised (57; 74).

One common subject that most of the aforementioned studies touch upon is the ethical implications of using such models. There is definitely bias present in the data training set which can cause various degrees of *hallucinations* of such language model. Lund et al.(56) and Meyer et al(61) agree that the design and purpose of such specialized LLMs play a significant role in addressing the biases. It is difficult to define metrics that assess the level of bias in the training data and in the model outputs. Moreover, further research should evaluate through the human-computer interaction lenses since the users seem to manifest an unfounded high degree of trust into such CRS. LLMs such as ChatGPT would benefit from the ability to distinguish between queries that it can safely handle and those that should be redirected to human experts (eg. health issues)(29). Cai et al. (12) suggest taking into account users' personality traits, particularly conscientiousness, as an important step in designing recommender systems. For individuals with higher conscientiousness, who prefer to thoroughly evaluate all aspects before deciding, a *Mixed-Initiate system* that allows both user-initiated and system-initiated interactions is preferable. Therefore, ethical implications should be considered from the design phase, highlighting the user interaction flow based on well-defined user requirements.

2.5 Evaluation metrics and methodologies for abstract generation

Explanation of evaluation metrics commonly used to assess the quality of generated text, such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) and human evaluation. An ongoing discussion of the appropriate evaluation methodologies to measure the effectiveness of your language model conversational recommender is caused by lack of standardized evaluation framework (49; 69). The most often benchmarking methods are using F1 score, Accuracy and Precision, ROGUE and BLEU along several variations (87) of them as metrics designed to compare the generated text against reference (ground truth) text and provide a numerical measure of similarity or overlap. It has been shown that such metrics have low correlation with human judgments. Therefore tailored metrics based on LLMs are being researched with the goal of facing the decision-making and cohesiveness issues common in LMs (53; 40).

For the human evaluation, there are several ways to conduct it in order to measure the conversational level of a LM. One of the most common ones is through user experiments, where the user is asked to converse freely with the conversational system and later rate the answers and conversational flow (27; 55). A popular metric of assessing the ability to respond in a humane manner to the user is the popular SSA (Sensibleness and Specificity Average) popularized by Adiwardana et al., which boils down to specific questions regarding the interaction. Another human evaluation method is to have humans annotate the answers of the system as a measurement for factuality and comprehensiveness (102). This can be done through diverse approaches, such as Ranking/Preference Judgments, quality ranking, task-specific metrics or preference comparison.

Bhandari et al. (7) point out that the choice of metrics depends not only on different tasks but also on different datasets and application scenarios. A mix of quantitative and qualitative metrics seems to be appropriate in evaluating CRS from different perspectives, which can further contribute towards gaining valuable technical and user-related insights.

2.6 Research structure

Design science involves the study and creation of artifacts explicitly intended for human use, as defined by Johannesson (36). It aligns well with the software product development cycle and incorporates scientific methodology to rigorously define, explain, motivate, and design the subject of study. In this work, we adopt design science as a framework to describe the process undertaken, as it naturally fits alongside Agile Software Development processes. Previous studies by Adikari et al. (2) show that design science's iterative approach between design and development better suits Agile requirements engineering than alternatives, which may face challenges with unforeseen requirements due to the just-in-time nature.

2.7 Iris.AI background

Existing language models, despite their power, may not fully grasp the distinctive features of scientific articles, leading to suboptimal information retrieval and abstract generation. LLMs are often fall short on providing in-depth nuances, processing out-of-vocabulary words and dealing with the challenges of factuality. Iris.AI, an innovative company that specializes in artificial intelligence and natural language processing for scientific research. Founded in 2015, Iris.AI is on a mission to revolutionize how researchers access and process scientific knowledge. The company's flagship product, the *Researcher Workspace* platform, is a cutting-edge AI-powered research assistant that aids researchers in navigating and comprehending vast amounts of scientific literature efficiently. Researchers can interact with the platform through natural language queries, allowing them to find relevant scientific articles, patents, related work and insights quickly. Some of the main models used in the platform besides the Smart language model are a topic model, an information extraction model, a word pieces model and an entity recognition model. The users and collaborators consist mostly of academic institutions, R&D companies and research organizations.

With Iris.AI's user-centric approach and focus on scientific applications, the company is reshaping the landscape of knowledge dissemination and research productivity. By simplifying access to scholarly literature and enabling researchers to efficiently obtain insights, Iris.AI plays a vital role in accelerating scientific discovery and promoting collaborative knowledge exchange among researchers worldwide. To further improve the usability of the platform and increase user's satisfaction, the proposed master thesis seeks to explore and implementation of prompt engineering techniques for the Iris.ai conversational recommender system. By refining the language model's prompt generation strategies and utilizing domain-specific pre-training, the study aims to create an innovative conversational recommender capable of assisting researchers in obtaining comprehensive and concise scientific abstracts as well as helping through conversation filtering the most relevant articles for the user search.

Chapter 3

Methodology

This section presents the methodology employed to achieve the goal of developing a language model capable of generating a 300-500 word scientific research related abstract based on the conversation with the user with a specific focus on scientific articles as well as acting as search query filter. By leveraging the advancements in Natural Language Processing (NLP) and Human-Computer Interaction (HCI), this research aims to address the need for automated text generation of scientific literature, facilitating efficient knowledge extraction and dissemination. In order to capture relevant pieces of the big picture, both qualitative and quantitative method are intermittently used in preliminary research, design and evaluation. Inspired by Walsh's (90) applications of *grounded theory*(GT) (28) in Information Systems, we consider that opting for an actionable solution rather than a prescriptive one can still fill in significant research gaps even if the research approach follows more of an exploratory paradigm. In this section, the intended methods will be discussed for each relevant phase of the research together with later developments, which will highlight the differences between the planned research and the actual executed one.

Main goals of this research:

- Create a requirements wishlist for CRS system that uses prompting techniques
- Have user evaluation on two conditions (one Q&A, one conversational) in order to figure out the target group's needs and behaviour
- Benchmark hard and soft prompts to assess the models' performance

3.1 Design approach in research methodology

The research strategy utilized in this endeavor is essential for comprehending and addressing the research questions. In the following, we will describe the significant steps of our chosen strategy. Design science research can be defined in various manners, yet it is widely acknowledged as encompassing the investigation of an artefact's design and development to gain insights into the process of creating the said artefact(68). In design methodology, an artefact refers to a tangible or intangible object, product, or prototype that is created as part of the design process(8). The artifact plays a crucial role in design science research and other design methodologies, as it provides a means to assess the feasibility, functionality, and effectiveness of the proposed design solutions. The ultimate objective is to develop a comprehensive solution tailored to customer needs for various use cases. Simon(84)[p. 55] illustrated accurately the relevance of design science in research methodology by saying "Whereas natural sciences and social sciences try to understand reality, design science attempts to create things that serve human purposes". During the investigation phase, which aims to acquire knowledge, there is a possibility of challenging the initially envisioned approach through later refinements of the prototype.

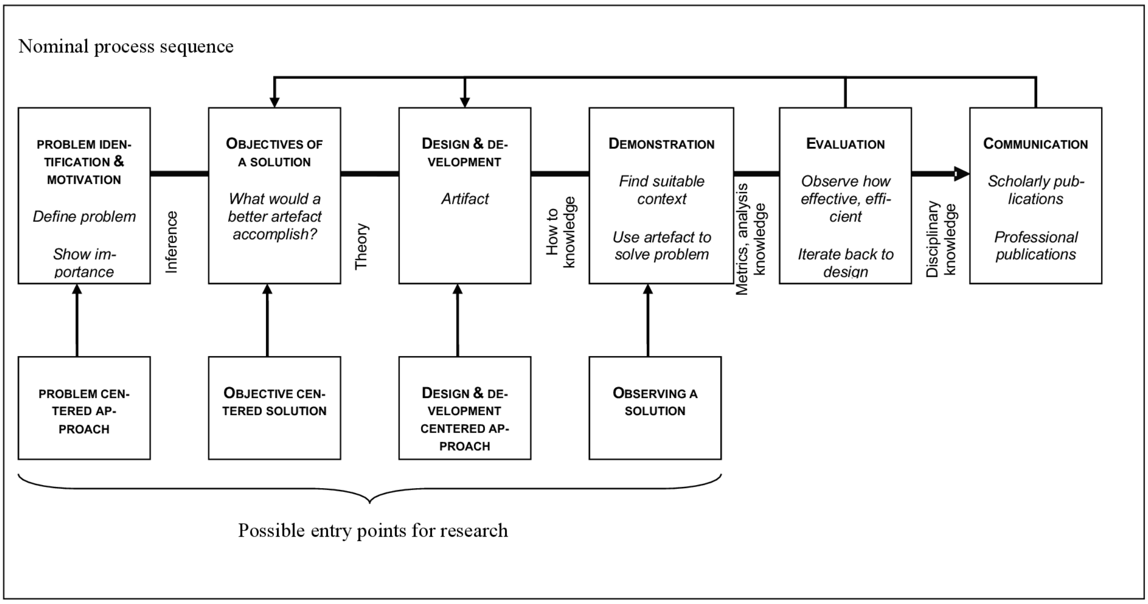

Figure 3.1: Design science research framework, Peffers et al.(2007)(68)

As one can follow in Figure 3.1, the major steps in design science are: explaining the problem, defining the requirements, designing and developing of an artefact, demonstrating the artefact and evaluation of that artefact. According to the evaluation results or, perhaps, the impossibility of demonstrating an artefact, previous steps, and their outcomes have been revisited and updated, which is in line with the iterative approach of design science. In our case, we have had two conceptual loops, the first resulting in the redefinition of the problem by adapting the hard and soft prompting to two different tasks and the second being an iterative development-evaluation process. Proofs of concept have played a pivotal role throughout the implementation process, guiding decision-making at each step and ultimately becoming integrated into the final feature implementation. Evaluating interface compatibility between the platform and the proposed solution, updating the requirements for version compatibility with the API and libraries used, and focusing on a user-centered design approach have been early testing processes that aided later parts of the development and demonstration process. In Chapter 5, all the latest developments and deviations from the methodology will be discussed at large.

Measurements

Qualitative answers definition

There is no standard definition of what a qualitative answer is. This is a known issue when it comes to the answers of LLMs. To define the meaning of more qualitative answers from the user, there has been settled upon a set of features to assess the quality, such as robustness of the answer, follow-up answer match, correctness and accuracy with regards to the question and numbers of conversational topics found in an answer (4; 62). To fairly compare the two results (hard prompts and soft prompts), we will use the expertise and knowledge of the experts to assess the quality and usefulness of the answers, mainly for two reasons: 1) the purpose of this research is to contribute to further developments of scientific domain CRS and text generative AI, and 2) measure the helpfulness of such system for onboarding the user.

3.2 Data collection

Data collection for the prompt tuning model was conducted through various methods to gather essential insights and requirements from different stakeholders. A mix of quantitative and qualitative methods have been employed in both the design and evaluation phases of the research. Based on Walsh's (90) comparison of different grounded theory research designs, we designed the Figure 3.2 that reflects the approach considered the best in helping us not stray away from the substantive area of investigation.

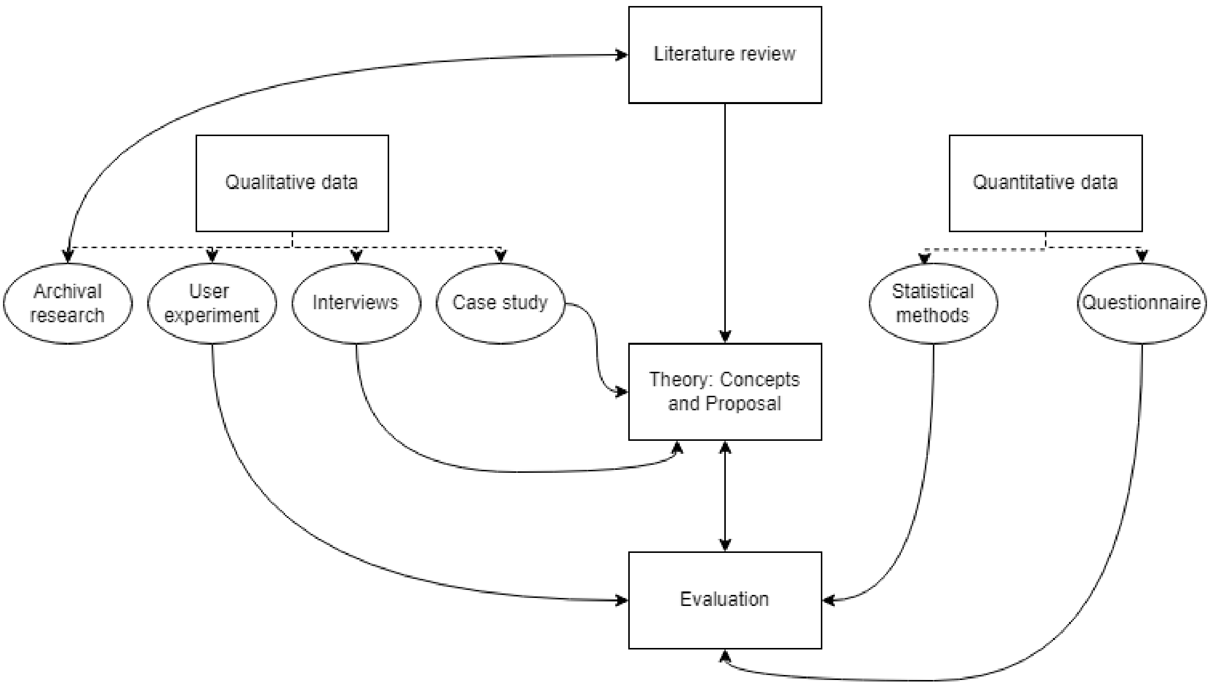

Figure 3.2: Overview research methodology flow

Using internal information is a good source for the problem's requirements definition and identifying how businesses encounter the problem and how a solution might address their needs. However, relying on publicly available information for technical documentation, decision-making, and the development process offers greater reliability as it eliminates individual biases. Moreover, adhering to the ethos of open-source technology infrastructure aligns with our principles, as open-source systems typically have faster patching times for vulnerabilities compared to closed-source alternatives. According to Google's Project Zero reports from 2019 to 2021 (78), open-source projects generally exhibit quicker response times. Additionally, The following methods were employed:

Preliminary Research

In the initial phase, interviews were conducted with employees and users to understand their requirements for the conversational recommender system. The interviews focused on capturing user experiences and identifying any issues or limitations with the current platform. The decision to use the following methods was made to align with the Agile product development process and strike a balance between meeting stakeholder needs and ensuring technical feasibility. The design and development process outlined in this paper relies on several pillars, each with its own data collection method. Throughout the iterative process, these pillars have influenced decisions regarding whether to continue with the current path, make adjustments to the objectives, or pivot the approach accordingly. This section will include pre-design research methods, ranging from interviews to proof of concept iterations.

Archival research: To begin, the current research on prompting techniques is still considered a fresh topic within the industry, given that large language models and their integration with popular applications are a current trend in the field of AI. *The literature analysis* helped offer clues about recent research developments of the Conversational Recommender system, which large language models are potential candidates for this research, the advantages and drawbacks of using prompt techniques and evaluation frameworks and metrics in measuring not only the model's efficiency but also the user's satisfaction and ease of use. By following the state in the industry, we have concluded that some of the existing solutions are partially developed or lacking in terms of accuracy and user evaluations. By comparing the performance of the models with hard and soft prompts, the study aims to identify the most effective prompt engineering approach for navigating within the context of a scientific literature database.

Interviews: A core method for *qualitative data collection* has been represented by direct stakeholder interviews, in which employees from both technical and business departments spoke on behalf of several users about usability problems, possible improvements of the platform, user experience flow and remarks on the proposed solution. Some of the needs partially overlapped and can be found in the requirements list, accompanied by the agreed use cases guided both the iterative process and the final target workflows. We concluded human-like generative answers and search query function are the main focus for the design and implementation and, as a result, factuality in a scientific abstract generation took a secondary place. The sessions were conducted in an informal online format involving various actors (apart from the development team and customer representatives). There were two types of sessions: user-focused back-and-forth discussions and proposal presentations. In total, there have been four one-time interviews with people from marketing, sales, organizational and front-end teams, as well as numerous sessions with the head of Research team lead and CTO.

In the user-focused discussions, Iris.ai business and technical team mentioned current users' remarks, and after receiving feedback and outside deliberations, they took shape as a common accord proposal. The proposal has not specified either architectural models or implementation details but rather represented a plan of action for the future versions of the platform where more user centered features such as the conversational recommender become a priority. We committed to the project's timeline, to investigate use case feasibility, and, if a positive result arises, to implement the desired deliverables.

Iterative Development and Feedback: *Proof-of-concept (PoC)* experiments and fast iterative processes have directed multiple positive decisions and helped in weeding out out-of-scope details. One such first PoC showed that there should be two separate experiments for hard and soft prompting rather than one that integrates both parts. Through two more proofs-of-concept 3.3 that successfully accomplish the tasks of generating a scientific abstract based on a conversation and acting as a search query filter to narrow down the results shown in the database, respectively, it has been shown that the plan to solve the problem can be accomplished and that the previous decision of focusing solely on the scientific abstract generation can been expanded upon. As a result, the architecture design has been finalized and committed to, and the PoCs have been refactored to become modules in the final workflow.

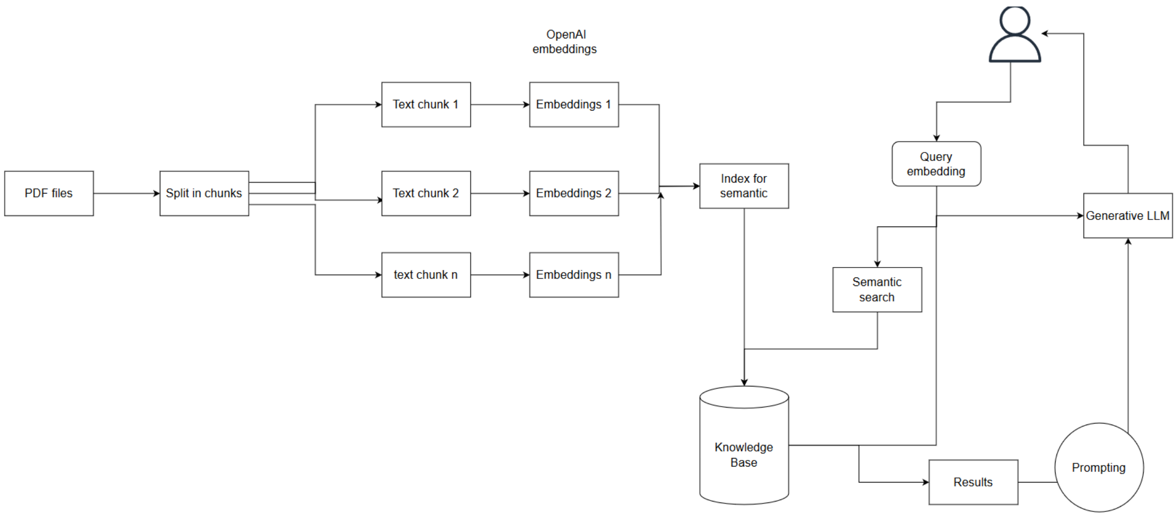

Figure 3.3: PoC diagram flow for conversing based on PDF

By employing these data collection methods, we were able to gain a comprehensive understanding of user needs, refine the prompt tuning model, and ensure that the conversational recommender system effectively addressed user preferences and challenges.

Experimental design

Three methodologies were used to evaluate the interactions with the CRS: the SASSI and the SSA augmented questionnaires, as well as benchmarking the performance of the soft prompting model. By doing so, the study makes use of both quantitative and qualitative methodologies to compare and evaluate accurately the soft and hard prompting techniques.

SASSI questionnaire: The SASSI questionnaire is used for quantitative data collection and consists of a self-report measure of the UX of the evaluated system. The questionnaire focuses on measuring the general speech-system usability as well as the user's experience. It is usually used for evaluating speech-oriented systems, such as conversational speech agents. However, given the similarities in requirements of speech and

Table 3.1: SASSI - Evaluation of System Characteristics

Items	**# of Sub Items**
System Response Accuracy	9
Likeability	8
Cognitive Demand	5
Annoyance	5
Habitability	4
Speed	2

generated conversational text, we adapted SASSI items to fit this research. The only words changed were where "speech" or "speak" was present in the statements, we added a "write". The choice for SASSI was based on a comparison between multiple assessment tools for user experience (UX) in conversational interfaces(39), where SASSI fit the best with the context of this research due to measuring dimensions such as Enjoyment/Fun, and Frustration, but also as an overall combination of all other UX dimensions when interacting with a CRS.

In total, the SASSI consists of 6 categories likeability, speed, habitability, cognitive demand of the user, system response, and habitability. Each statement had to be rated on a Likert scale from 1 to 6, where 1 means "Strongly disagree" and 6 is "Absolutely agree". An open-ended question was added at the end where the participant had the freedom to add any remarks/suggestions of improvement regarding the system, which helped in gathering the user's opinions on all of the three methods.

SSA augmented questionnaire: To measure the quality of a response given a context, the conversation between the agent and the students will be evaluated by three people who are experts in their field and work at Iris.AI. In this study, we are going to refer to them as experts. Inspired by the work of Adiwardana et al. which used the sensibleness and specificity average (SSA) metric(55) calculated from the sensibleness and specificity scores we tested three variables: sensibleness, specificity, and utility.

The first score, sensibleness, measures if an utterance fits the general context, evaluating whether a model's response is appropriate in light of the surrounding information and does not conflict with anything expressed until that moment. Humans frequently take this fundamental criterion for communication for granted. At the same time, generative models frequently struggle to meet this requirement, escalating in unintentionally rewarding models for playing it safe by consistently producing brief, generic, and uninteresting responses when sensibleness alone is utilized to judge models. The GenericBot algorithm(55), which always responds to queries with "I don't know" and to statements with "Ok," achieves a sensibleness score of 70%, even exceeding certain sophisticated dialogue models.

A response's specificity to a particular context is evaluated using the second score, specificity. A user would say, "I adore Florence," and the model might reply, "Me too." In this case, the model would receive a score of 0 for specificity because this response could be applied to a variety of situations. "Me too. I love Santa Maria del Fiore Cathedral," would receive a score of 1. According to Adiwardana et al., Meena closes the gap to the human performance average in the SSA metric (55).

However, when applying the SSA metric to questionnaire answers, sensibleness and specificity are insufficient to assess a dialogue's quality. For instance, a logical and pertinent solution to the question "Could you generate me an abstract for a paper on OCR wtih Graph Neural Network?" would be creating an actual representative abstract for the given topic. The response which would contain details about the research question, chosen methodology and fictive results would be a deeper and more gratifying alternative. This is the reason for introducing the usability metric in the questionnaire, making it more similar to the Sensibleness, Specificity, Interestingness (SSI) metric (86) used by Thoppilan et al. where they introduced "Interestingness" as a third Boolean variable to measure whether an utterance is perceived as likely to "catch someone's attention" or "arouse their curiosity", or if it is unexpected, witty, or insightful.

For each of the three metrics, they check a box if an utterance was rated sensible, specific, or useful, resulting in three Boolean values. While the score has been thought for evaluating language models for dialogue application, this study calculates it for the system's answers. Sensibleness contributes to understanding whether the system's answers fit the context or if a question is misinterpreted. If an answer is labelled as sensible, we further ask the experts to determine whether it is specific to the given context or could be generally used for different situations. This helps us identify if the system's answers are too generic to produce valuable feedback. Usability's objective is to measure the quality of a system's response by asking experts if the answer provides good feedback for the course's improvement.

Two open-ended questions are found at the end of the questionnaire to assess the difference between the two conditions and to provide the opportunity to present eventual remarks, allowing the collection of qualitative insights.

- Could you describe if one or more of these methods helps in achieving the task of generating an abstract and/or information retrieval in the given domain?
- Are there any remarks that you would like to point out?

Hard prompts

Iterative refinement of prompts allows for a data-driven approach to prompt design, leveraging feedback from multiple conversations to improve the prompt effectiveness. It heavily relies on the few-shot approach. By systematically assessing the generated output for coherence, informativeness, and relevance, prompt refinement ensures that the language model learns to capture the most salient aspects of the scientific articles. However, this constitutes a subjective pre-assessment in the quality of prompts based on the researcher's biased evaluation. Thus, the refinement of hard prompts is subjected to biased iterative refinement, rather than relying on multiple user evaluations. For this research, 100 system-user outputs have been refined until the hard prompts match the desired output for both tasks.

The construction of the hard prompt had three main parts:

- Prompt components: deciding which components are required for the prompt structure. The role was the first component, followed by context, task instruction and lastly, few-shot examples.
- Domain-Specific Considerations: taking into consideration the target audience of the end product, mainly researchers and academic employees, the tonality should

consist of semi-formal, highly accurate and in-depth details, including specific abbreviations and symbols used in the domain and adhere to the IMRAD standard of abstract generation structure. This point brought into discussion the training set of the language model.

- Prompt formulation and variation: to obtain the most concise and coherent output, several edits had been made in the wording, sentence structure and placement of the text.

The iterations of the hard prompt included the addition of extending the context for more in-depth answers, defining the role of a research assistant, implementing conversation memory, and few-shot examples of always ending the system answer in a question (eg: *"Should I generate an abstract or continue the conversation?*) to support turn taking in the conversation, increasing default token limit size of the generated answer, styling the manner of speech to be conversational friendly instead of question and answering format, creating a chain of thought in the prompt to minimalize unfaithful and incorrect answer, link retrieval from scientific sources. The final hard prompt, although apparently long and complex, yielded the most constant and accurate behavior of the model (see Table A.4.

Soft prompts

Soft prompts are acquired through backpropagation, enabling their customization to incorporate signals from various labeled examples. Following the explanation of the utilization of soft prompts with a frozen model from Lester et al. (42) we will explain how prompt tuning is embedded into in LLM architecture. Instead of modeling classification as $P(Y|X)$, where X represents a series of tokens and y is a single class label, we now adopt a conditional generation approach, represented by $P(Y|X)$, where Y is a sequence of tokens representing a class label.

In order to gain a more comprehensive understanding of the mechanics involved in this process, the first stage entails the embedding of a series of n tokens, denoted as $\{x_0, x_1, ..., x_n\}$. This embedding process results in the construction of a matrix $X_e \in R^{n\times e}$, where the symbol e represents the dimension of the embedding space. The manifestation of our soft prompts is represented by a parameter $P_e \in R^{p\times e}$, where p denotes the length of the prompt. The combination of our given prompt and the included input results in the creation of a concatenated matrix represented as $[P_e; X_e] \in R^{(p+n)\times e}$.

Prompt tuning will be used as part soft prompt technique in both tasks. In difference to Fine Tuning no changes are required in the pretrained model weights (parameters). This makes a significant difference since changing the underlying model is not a cost and time efficient approach. The model remains 'frozen' while being adjusted to the domain and user-specific requirements.

3.3 Model

Through this methodology, the study seeks to provide valuable insights into the effectiveness of hard and soft prompts in combination with different language models for scientific abstract generation. In this study, the methodology encompasses these two key components, adapting them to the unique requirements of **carrying a conversation with the**

goal of generating scientific abstracts and **identifying relevant articles related to search query based on conversation with the user**.

On one hand, the significance of prompt engineering lies in its ability to guide the language model's generation process, effectively influencing the quality and coherence of the system's output. By designing conversation prompts that encapsulate the crucial elements of a conversation geared towards exploring scientific concepts, the latest research, and a variety of experiments, including key findings, methodology, and significance, the language model can be directed to generate answers that encapsulate the essence of the desired research based on factual knowledge.

On the other hand, prompt-tuning is an efficient, low-cost way of adapting an AI foundation model to new downstream tasks without retraining the model and updating its weights. For this task, based on literature review and similar models, we chose three PTM which would suit this research. SciBert(5), a well-recognized language model derived from BERT(18), has been designed to possess exceptional capabilities in interpreting and extracting complex information from scientific publications. As a result, it has become a valuable resource for effectively understanding the subtle language commonly seen in academic and technical discussions. In contrast, GPT-2 (72) shows a strong capacity for generating content, enabling the production of coherent and contextually appropriate system replies. This attribute is crucial in developing conversational recommendation systems that aim to be effective. The Falcon (79), renowned for its exceptional efficiency and remarkable speed, emerged as a highly appealing option for guaranteeing instantaneous responsiveness in user interactions. The three pre-trained models have been chosen based on their versatility, GPT2 and Falcon falling into open-domain category of LLM while SciBert has been specifically pre-trained on research articles. The public dataset used for all the models consists of abstracts of scientific papers.

3.4 Ethical concerns

To make sure this research is ethically viable, an ethical request has been documented and has been approved by the Ethical Committee of the Faculty of Electrical Engineering Mathematics and Computer Science (EEMCS) from the University of Twente and checked against the ethical requirements of Aalto University. Participants participated voluntarily in this study. The participants could withdraw before, during or immediately after the experiment. Moreover, there were no risks involved in participation. With respect to the system, having a conversation with a CRS does not lead to suggestive prompts. Instead, the CRS responds to the provided answers of the participants. At last, at all times, the identity of the participants, including their gathered data, who have participated in this study, will remain anonymous. The conversation between the participant and the agent has been recorded in a log file (with permission as stated in the consent form (see B). Only participants who agreed to this, were included in the research. The files have only been stored in the institutional-affiliated cloud storage of the researcher for a maximum of two weeks. After this time had passed, the files are deleted.

However, designing and developing language model conversational systems raises significant ethical concerns and challenges related to biases. Language models are trained on vast amounts of data from the internet, which can inadvertently contain biased content reflecting societal prejudices. As a result, the language model may generate responses that perpetuate stereotypes, offensive language, or harmful ideologies. Moreover, biases can arise from imbalanced data, leading the system to favor certain perspectives or demographics. Developers must be diligent in mitigating these biases and ensuring fairness and inclusivity in the system's responses. In the future, more transparent and robust testing procedures should be implemented to identify and address biased outputs. Additionally, concerns about privacy and data security arise when conversational systems store user interactions. Safeguarding personal information and ensuring user consent for data usage are critical aspects of ethical development and will be further discussed with Iris.AI regarding the implementation of such a system. Ethical guidelines and responsible use policies are essential to safeguard against these risks and promote ethical behavior in the field of language model conversational systems and will be prioritized in the design of the artefact.

AI generated topics

In the course of our investigation, we employ ChatGPT(65) as a tool for generating scientific study subjects. The expansion of our study's scope and variety presents ethical considerations. We recognize the possibility of biases in content generated by artificial intelligence and pledge to undertake a comprehensive examination and improvement procedure to guarantee impartiality and pertinence. The importance of transparency is emphasized in our approach, as we clearly acknowledge that certain topics are derived from AI sources. This serves to underscore our commitment to the responsible utilization of AI and the implementation of unbiased research methodologies. Throughout the course of our investigation, we consistently uphold an ethical standpoint, ensuring that the contributions made by artificial intelligence serve to strengthen the integrity and inclusivity of our scholarly inquiry.

Chapter 4

Collected requirements

This chapter includes a curated wishlist of criteria that resulted from interviews with IRIS AI, and a review of the relevant literature. This list is beneficial in the sense that it may be applied to both the academic and the industrial domains as a set of guidelines for future research and experiments. It is essential to keep in mind that the items on this wishlist are not solutions but rather goals that offer a vision for the future of conversational recommender systems implementation. The requirements that are technical, user-focused, and functional together with the scenarios relevant to our topic of research provide a better understanding of how to design the system and user evaluations. Even though not all of these have been put into practice in this research, they nonetheless serve as a helpful guide for conversational recommender systems and prompting techniques research, with the goal of ensuring that these systems can adapt to the changing requirements of academia as well as industry. At the end of this chapter, we will highlight which requirements are satisfied by this research.

4.1 Problem formulation

The main focus of this research is to improve the user adoption process and address the limitations of the existing Iris.AI platform, with regards to the *Explore tool* feature. Currently, users are required to provide a specific research question and a detailed problem description or article abstract, which narrows down the search for relevant scientific papers. However, this approach assumes that users already know their research topic and are seeking articles related to it, excluding those in the exploration phase.

To overcome this limitation and make the platform more user-friendly, we aim to develop a prompt-tuned conversational recommender. The designed system will allow users to explore and discover research topics without needing a predefined research question, making the first interaction process smoother and inclusive. Additionally, we want to optimize the process of generating abstracts and filtering relevant articles. After generating the abstract and forming the article database, the conversational recommender will rank and retrieve the most relevant articles based on the user's input. This will improve the efficiency of information retrieval and help the user in getting a glimpse of the dataset, thus guiding the user in deciding if the dataset is the right one for their research.

By focusing on these challenges, we hope to leverage prompt engineering strategies in the language model conversational recommender to enhance its accuracy, relevance, and user satisfaction. Our goal is to better support researchers in accessing relevant scientific papers and improving their productivity during the research process.

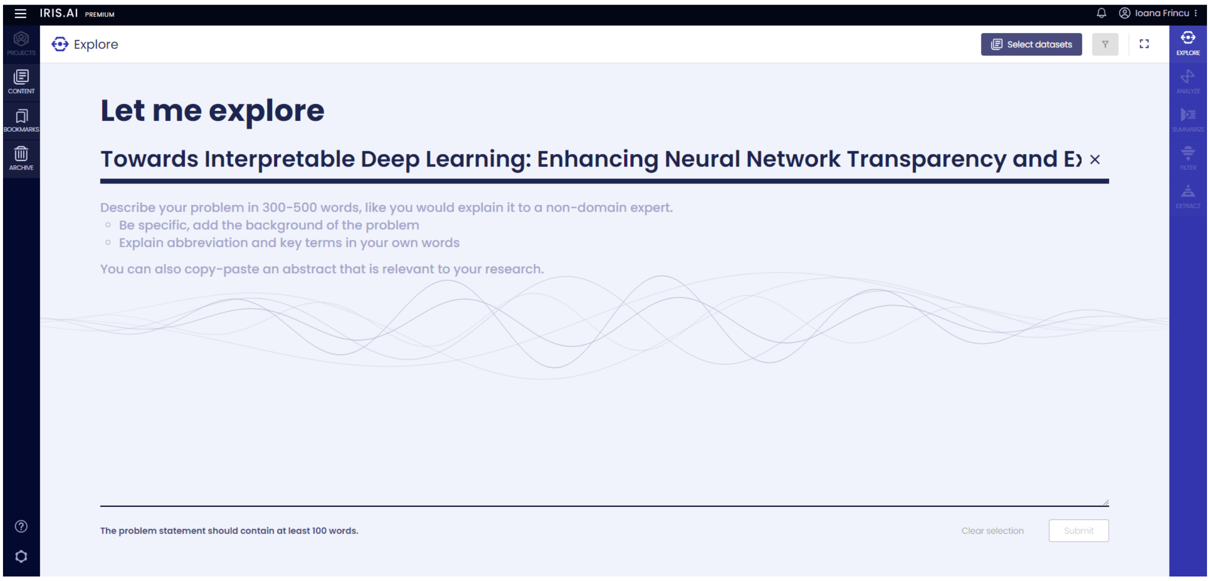

Figure 4.1: Explore tool start

4.2 Requirements

To create an effective prompt-tuned language model conversational recommender for scientific abstract generation, we need to address a set of technical, functional, and user requirements. These requirements lay the foundation for building a robust and user-friendly system that caters to researchers' diverse information needs. All requirements were formulated after interviewing the employees of Iris and from literature review which helped combine different perspectives into specific requirements. This is a wishlist rather than of the exact requirements which were implemented in this research, however we will mention what we use in this research and justify why we believe it is an adequate baseline for future studies.

Technical Requirements

- **Language Model Configuration:** The language models must be carefully selected and configured to support soft prompting and hard prompting. It should be capable of handling conversational interactions effectively and generating contextually relevant outputs. It should be a pre-trained model that can effectively cut down on time and resources spent on training them. One of the models should be pre-trained on scientific databases which allows for a comparison in performance between it and other general PTM. The models should be able to perform multiple tasks, such as generating text and retrieving information on the sources of the generated output. Exclusively open-source models were primarily chosen due to the advantage of increasing scientific replicability.

- **Prompting Techniques:** The language models and frameworks should be equipped to incorporate conversational chains which incorporate memory buffers, prompt templates functions, soft prompting through freezing the model weights, and updating the parameters of the prompt. Both hard and soft prompts are implemented through open-source LangChain framework which enables access to multiple prompting functions and NLP libraries.

- **Natural Language Processing (NLP) Capabilities:** The system must have strong NLP capabilities to accurately understand and process user input, ensuring that the generated outputs align with the user's intent and context. Efficient tokenization, contextual understanding, and a capacity for generating coherent, contextually appropriate text are vital. Thus, all models are versatile NLP pre-trained models.

- **Database Management:** Since the model are pretrained, the only database management will consist of embedding Chroma database or simmilar vector databases for both and soft prompts.

- **Processing power:** Running the model locally in Python demands substantial computational resources. The system must have access to a high-performance CPU or GPU to handle the complex calculations involved in training and inference efficiently. Google Collab, AWS or other cloud-based platform are often essential step to make sure the performance of the models is steady.

- **Environment:** Python 3.9

- **User interface:** The *Streamlit* library was used for developing a quick UI for this research and it is an open source easy to use application for CRS.

Functional Requirements

- **Abstract Generation:** The primary function of the conversational recommender is to generate high-quality abstracts for scientific papers based on user input. The abstracts should contain context and be formulated based on scientific papers' structure. Also, they should be relevant to the user's query, either given directly to one query regarding to the topic or based on the attached papers from the user and a conversation.

- **Prompt-tuned Interaction:** The system should facilitate natural, prompt-tuned interactions with users, allowing them to explore and discover research topics without needing a predefined research question. The prompt should handle specific cases, such as out-of-vocabulary words, terminating the conversation, and responding adequately when unable to generate the most appropriate answer.

- **Filtering Relevant Articles:** After generating the abstract and accessing the related articles to the topic, the system must effectively filter and rank the most relevant articles based on the user's input, ensuring efficient information retrieval.

- **Contextual Understanding:** The system should demonstrate a deep understanding of the user's context and intent during the conversation to generate accurate and contextually relevant abstracts. The system should be able to process the user's query correctly and avoid mismatching the intention. The system should also process correctly the input documents given to it which add new layers of knowledge about niche topics. If the user wants to continue the conversation, the system must be able to give context related information and not directly generate the abstract.

Otherwise, the system should just generate the abstract directly based on the given query.

- **Error Handling:** Robust error handling mechanisms are necessary to ensure the system can gracefully handle unexpected inputs and respond with informative error messages. If the system does not know about the topic, then it should specify "I don't know" rather than generating false answers. For each error case handling, the system should react accordingly.

User Requirements

- **Ease of Use:** The conversational recommender should include a user interface that is straightforward and easy to navigate, enabling researchers to communicate seamlessly and access abstracts without encountering any technical obstacles. For design should be intuitive and minimalistic.

- **Inclusivity:** It is imperative that the system is designed to be inclusive and accessible to researchers at all phases of the research process, including those in the exploratory phase who have not yet formulated a carefully defined research question. Besides researchers, graduate students and other academic or research staff should be able to use the CRS within its intended goal. Thus, the system presents both options of having a conversation with or without papers added as user input or to directly generate the abstract.

- **Accuracy and Relevance:** Users anticipate abstracts that exhibit a high degree of precision and contextual pertinence, thereby aligning with their information requirements and guaranteeing the generation of meaningful outcomes by the system.

- **Prompt Customization**: For researchers, it is crucial to have the ability to customize prompts and adjust the system's responses to match their preferences and specific research inquiries. This should happen during the conversation through the user's input. Enabling the system to discuss in-depth topics can be done by uploading papers that the model can be trained on in the hard prompting prototype.

- **Real-time Response:** To improve user experience and productivity, the system should provide prompt responses to user queries, generating abstracts in real time.

Our goal is to create an efficient and user-centric language model conversational recommender that supports researchers in accessing relevant scientific papers and improving productivity during the research process. By meeting some of these technical, functional, and user requirements, we aim to achieve this objective.

4.3 Use cases

The system seeks to improve researchers' access to scientific papers, increase their productivity, and facilitate the seamless exploration of research topics through its diverse capabilities. There are several use cases that require to be better defined which leads to a smoother implementation process. Using the design principles as a guide, the presentation of use cases helps distinguish between the main issues this research seeks to address and possible out-of-scope applications of the proposed system (41). In addition, it facilitates the discovery of exceptions and outliers in use cases, which can provide insights for future iterations of system design.

Use Case 1: Initial Query Expansion

Objective: To demonstrate how soft and hard prompts can be used to improve the user's initial query and the relevance of recommended scientific literature.

Scenario: Consider a scenario where a user initiates a search with a vague query such as "cancer research." In such cases, after giving an initial answer to some notable improvements in cancer research, the system can be used to extract more context from the user. For instance, the user could be asked to clarify their interests by specifying their search in a more specific manner, such as "I'm interested in recent advances in breast cancer research." Alternatively, the CRS can be used, which demands a specific query structure from the user. For example, the user could be asked to provide specific keywords related to their research on cancer. This approach is less flexible and may not be suitable for users who are unsure about the specifics of their query.

Benefits: Soft and hard prompts are two different approaches used in generating recommendations. A soft prompt allows for a more natural conversation, while a hard prompt enforces a structured approach. The soft prompt approach aims to create a more conversational experience between the user and the system. It allows the user to interact with the system in a more natural way, without feeling constrained by a rigid structure.

On the other hand, the hard prompt approach enforces a more structured conversation between the user and the system. It is designed to help the system generate more precise and accurate recommendations by asking specific questions that help the system understand the user's preferences better. Thus adding two options of the hard prompt prototype, one where there is a conversation and one where the abstract is generated immediately based on the user's input, caters to the diverse preferences of the user.

The impact of both approaches on the quality of literature recommendations and user satisfaction is significant. In the context of the Iris.AI platform, it helps narrow down the search query and provides results with a higher relevance score. Besides, there are some hypotheses that should be accounted for when testing the system. While the soft prompt approach may allow for a more natural conversation, it may not always generate accurate abstracts that meet the user's needs. On the other hand, the hard prompt approach may be more accurate in generating abstracts, but it may not always provide a satisfactory user experience due to its rigid structure.

Therefore, it is important for recommendation systems to strike a balance between the two approaches to ensure a satisfactory user experience while also generating accurate abstracts.

Use Case 2: Filter and Refine Recommendations

Objective: To demonstrate how soft and hard prompts can be employed to filter and refine the list of recommended scientific articles based on user preferences.

Scenario: Suppose a user receives the abstract from the CRS, but they want to get inspired by the results based on actual publications. In this scenario, the user can be presented with actual links to the papers.

One option is that the links to the papers are provided at the end; let's say no more than three links. It provides a suggestion to the user to refine their search without being too pushy or demanding, but it lacks specificity, not linking precisely to a specific piece of text from the generated abstract. This can be achieved through hard prompting, where the model is asked to retrieve any relevant articles related to the generated text.

The other option is that each relevant piece of text is referenced from an article. This way, the generated abstract is built on scientific articles. However, this might impact the speed of the real-time response, and it would probably cost more computational power. This can be done in soft prompting where the model is already pre-trained on a comprehensive database filled with scientific articles.

Benefits: When it comes to finding literature that aligns with users' preferences, prompting techniques can be quite effective. However, the effectiveness of each technique may vary. For instance, one technique might be more successful than others in terms of user interaction with the system while others might hold more value from an information retrieval accuracy. By evaluating each prompting technique, we can determine which technique is the most efficient for assisting users in finding literature that aligns with their preferences. Additionally, we can also analyze how user interactions with the system differ based on the type of prompt used.

NOTE: This scenario has proven to be liable for hallucinations, where some of the generated links are incorrect, inaccessible and irrelevant to the topic (59). The implementation should take into consideration this possibility and report on such behavior from the system.

Use Case 3: Contextual Understanding

Objective: To highlight how soft and hard prompts contribute to the system's ability to understand the user's context and tailor recommendations accordingly.

Scenario: Dr. Maria R (fictional character), an environmental scientist, is embarking on a new research project aimed at understanding the impact of urbanization on local biodiversity. She has decided to use our conversational recommender system to identify relevant scientific articles for her study.

Dr. R initiates the conversation by saying, *"I'm interested in studying how urbanization affects local flora and fauna. Specifically, I want to explore the changes in biodiversity in urban areas over the past decade. My research will focus on bird populations, tree diversity, and the role of green spaces. Can you help me find relevant literature?"*

The soft prompt embedded in the system's conversational interface actively listens and extracts essential context from Dr. R's initial statement. It identifies her research problem, key objectives, and even recognizes relevant keywords such as *"urbanization," "biodiversity," "bird populations,"* and *"green spaces"* without requiring her to provide a structured list.

Benefits: The utilization of gentle cues in this particular situation presents numerous noteworthy benefits. Firstly, these features facilitate the extraction of the user's research context in a natural manner, effectively collecting the study problem, aims, and pertinent keywords without interruption. The reduction in cognitive load on the user is accompanied with an improvement in the system's capacity to provide recommendations that are closely aligned with the user's research focus. The level of user satisfaction is enhanced by the convenience and intuitiveness of the interface. Dr. R perceives that the system effectively understands her needs, resulting in time and effort savings during the research process. In general, the utilization of soft prompts contributes to a user-centric and contextually sensitive interaction, hence enhancing the system's ability to effectively cater to the information requirements of the researcher.

Present in this research

For our research, we chose to focus solely on generating the main abstracts and only incorporate some of the requirements. All technical requirements were met, only the first two functional requirements (abstract generation and prompt tuned interaction) were completely satisfied and only ease of use and inclusivity from user requirements were fulfilled.

The utilization of both prompt kinds can be strategically employed to enhance the user experience and enhance the quality of recommendations, hence enhancing the effectiveness of the conversational recommender system for scientific literature.

Conclusion

In the context of developing a scientific literature conversational recommender system, it is crucial to elucidate the practical use cases where the application of soft and hard prompts can yield distinct advantages. These use cases not only serve as compelling illustrations of the technology's potential but also provide empirical evidence to support the choice of prompt techniques within the system. Below, I present several use cases that can be explored and studied within this academic research framework:

Objective	Use Case Scenario	Description	Example
1.	**User Query Clarification (Soft Prompts)**	Soft prompts guide users to articulate their information needs more explicitly. Useful for vague or broad search intentions.	"Could you specify the subfield of AI or any particular aspect you are interested in?"
2.	**Structured Information Retrieval (Hard Prompts)**	Hard prompts gather structured or specific information. Employ predefined questions or prompts with well-defined response formats. Enhance accuracy and targeting of recommendations.	Gathering publication year, author name, or key terms for research articles.
3.	**User Feedback and Refinement (Combination of Soft and Hard Prompts)**	Combine soft and hard prompts to improve user experience and recommendation quality. Use soft prompts to gather feedback and preferences. Employ hard prompts to obtain specific details for fine-tuning recommendations.	Facilitate continuous improvement in recommendation quality.

Table 4.1: Use Case Scenarios for Soft and Hard Prompts in Conversational Recommender Systems.

Chapter 5

Proposed system

The following section will integrate the implementation of both the hard and soft prompts within one Conversational Recommender System. To avoid confusion in following the narrative, all subsections will individually address the development and implementation of both prompt categories.

5.1 Dataset

There were several tries on different datasets such as ArXiV, cite_rec and other datasets consisting of at least one feature being abstracts of scientific articles. However, there we several issues encountered when training the models on such large datasets, leading more to fine-tuning the models rather than prompt tuning them. Thus,

Hard Prompts

One of the two central components of our design revolves around prompt engineering, which allows for more interactive and dynamic conversations with the system. Only data used in the creation of the hard prompts was subjectively selected by the researchers as inspiration in prompt construction for prompt refinement for the scientific domain (102; 93). Besides the predefined prompt which was used for implementing context and assigning a role to the system, we treat the user's input as a prompt within the history of the conversation which contributes to the contextual learning abilities of the model. The history window is no longer than ten user inputs based on the pilot study interactions.

Soft Prompts

To lay the groundwork for our investigation, we chose the dataset RAFT of scientific papers abstracts encompassing a diverse range of topics and domains used by Liu et al. (47). With 50 entries, the dataset was considered appropriate for prompt tuning. The dataset is meticulously curated from reputable sources (PubMed and ArXiv articles), with two key features: the abstract resulting from a summarization task of a long article and the title itself. The acquired dataset serves as a rich resource to assess the performance of the CRS across a spectrum of subject areas, refining the final generated abstract by the CRS.

Data processing involved data cleaning, including any irrelevant or redundant information. Textual data is subjected to tokenization with the Autotokenizer, which pulls the matching tokenizer from HuggingFace depending on the selected model, along with minimal stemming and lemmatization processes to enhance text normalization.

The dataset is divided into appropriate subsets for training, validation, and testing purposes following the most commonly used 80:20 ratio. This division ensures that the CRS is trained on a representative portion of the data while maintaining the ability to evaluate its performance on unseen instances.

5.2 Model Selection and Configuration

Hard Prompts

In the context of model selection and configuration, it is imperative to consider various factors influencing the choice of the appropriate language model. In this study, the utilization of OpenAI's GPT-3 Turbo model initially presented a compelling option for generating responses to user prompts. However, when using open-source models such as SciBert and Falcon a critical aspect that emerged during the evaluation process was the model's responsiveness. When alternative models were assessed, a notable increase in response loading time, exceeding 14 seconds compared to GPT-3 Turbo, was observed. This latency impacted the overall user experience, prompting a strategic decision to streamline and enhance response time efficiency. Consequently, the final implementation of the system used only GPT-3 Turbo, seamlessly integrated into the Langchain framework. This selection not only upheld the quality of responses but also ensured a more fluid and user-friendly interaction.

From the LangChain framework, *PromptTemplate, ChatOpenAI* and *ConversationChain* functions were used for creating the conversational chain between user and system by including the prompt template mentioned in Chapter 3.

Soft prompts

As previously explained, there is a growing need to enhance and utilize more comprehensive models in the context of production-grade applications. The implementation of prompt tuning, part of Parameter-Efficient Fine Tuning (PEFT) umbrella, approaches arises as a solution to address this requirement, offering potential benefits in terms of cost and time effectiveness. The efficiency is attained through careful optimization of only the crucial and contextually relevant parameters inside the neural network structure.

Instead of fully fine-tuning separate models for each downstream task, you can utilize a single pre-trained model with frozen weights and then train and update a smaller set of prompt parameters. As part of this research prompting for casual modeling and model configuration, instead of fully fine-tuning separate models for each downstream task, I utilize a single pre-trained model with frozen weights and then train and update a smaller set of prompt parameters. This strategy becomes increasingly efficient as models grow in size, leading to enhanced results as the model parameters scale. Specifically, I apply prompt tuning to train three models: GPT-3, Falcon, and SciBERT, using the *systematic_review_inclusion* subset of the RAFT dataset.

In the context of model selection and setup, research was conducted to determine the most optimal language model to produce abstracts from conversational input. Three well-known models, including **SciBert**, **GPT-2**, and **Falcon**, underwent extensive training using a small and specialized dataset called RAFT (3). Each model had distinct strengths and qualities that corresponded to various facets of the task at hand.

PEFT config

To understand the implementation of soft prompts, we shall emphasize critical aspects of model training and configuration. In the provided code snippet, a configuration object named **peft_config** is instantiated for a natural language processing task associated with Prompt Tuning. This setup entails several vital configurations. The *task_type* argument is initially set to `TaskType.CAUSAL_LM`, signifying the engagement in Causal Language Modeling. This implies that the model generates text sequentially in response to given stimuli. The `prompt_tuning_init` parameter is configured as `PromptTuningInit.TEXT`, indicating that the prompt tuning process commences with text-based initialization. Subsequently, the `num_virtual_tokens` option is established as 8. Virtual tokens are employed to adjust the model's behavior during prompt customization, and the choice of eight virtual tokens demonstrates a nuanced control over the tuning process. The initial text prompt resides within the `prompt_tuning_init_text` parameter: `"Generate a 400-word scientific abstract based on the research topic."` This text prompt provides clear instructions to the model, elucidating the task it is expected to accomplish.

Tokenization

The correct padding token for sequence processing must be established, and we also need to know the maximum length (`max_length`) a tokenized label can be. Using `AutoTokenizer from_pretrained`, we begin the process by selecting a tokenizer based on the pre-trained model or path that has been provided. This ensures alignment with the chosen model for efficient handling of text data. Then, using the condition `tokenizer.pad_token_id` = *None*, we verify the presence of a designated padding token ID. When it is absent, we use the formula `tokenizer.pad_token_id` = `tokenizer.eos_token_id` to assign the end-of-sequence (eos) token ID as the padding token ID. In the preprocessing workflow, we first tokenize both the input text and labels. Then, for each example in a batch, we pad the labels using the tokenizer's (pad_token_id). Next, we combine the input text and labels into a unified (`model_inputs`) structure. We create distinct attention masks for both labels and (`model_inputs`). Finally, we iterate through each batch example again, padding the input IDs, labels, and attention masks to match the specified `max_length` and convert them into PyTorch tensors. .

Training

In the pursuit of advancing natural language processing, this research endeavors to train an optimized PeftModel. Initially, we initialize a base model employing the AutoModelForCausalLM class, pre-trained and fine-tuned for our specific task. Subsequently, utilizing the `get_peft_model()`function alongside a predefined configuration object named `peft_config`, we engineer a PeftModel. This novel model architecture offers training efficiency by fine-tuning only a fraction of its parameters while preserving the core pre-trained knowledge. The quantitative assessment of this efficiency is exposed as we print the ratio of trainable parameters to the total parameters, yielding an insightful "trainable" metric.

```
model = AutoModelForCausalLM.from_pretrained(model_name_or_path)
model = get_peft_model(model, peft_config)
print(model.print_trainable_parameters())
"trainable params: 6,144 ||
 all params: 124,445,952 ||
trainable: 0.004937083047908222
```

Following model construction, we set up key training components: an AdamW optimizer for updating model parameters and a learning rate scheduler for dynamically adjusting the learning rate. After establishing these fundamental components, the model is sent to the GPU for accelerated training. Each epoch in the training process contains a training loop as it progresses. The model is active in both training and evaluating modes for the duration of each epoch. We iterate through the dataset for the training phase, computing and backpropagating losses, and refining model parameters. Using a scheduler, we also carefully control the learning rate.

In conclusion, this study begins the training of a PeftModel, a method for enhancing model effectiveness in the field of text generation. A well-organized pipeline that includes model initialization, configuration, optimization, and evaluation serves as the foundation for our efforts.

5.3 Prompt Design

Designing the hard prompts was an iterative process, with plenty of trials and errors. Important to keep in mind that the decisions made in designing the hard prompts are subjected to our own biases when talking about the quality of the prompt. In this section we will explain the chain of thoughts in pinning down the most relevant design choices for constructing the right prompt for our tasks.

One of the useful sources for prompt design comes from a recent research from Zamfirescu (97). The study highlights the difficulty of non AI experts in designing prompts, one of the biggest issues being the opportunistic attitude rather than systematically one. When designing the hard prompts, we aimed to avoid design choices drawn from human social expectations such as polite language and maintaining a balance between examples and instructions.

Our system assumes the role of an "Assistant Researcher" who is tasked with the dual objectives of facilitating information retrieval while engaging in scientific discussions. We explain in detail the expectations on the output's comprehensiveness, which focuses on accurate and detailed scientifically grounded explanations. Defining the goal of the system was of utmost importance since it is more complex than just question and answer. Involving abstract generation, conversational text generation and source retrieval proved to be challenging. However, our system generates abstracts only when users request them, ensuring that it remains attentive to the ongoing conversation. To keep the conversation flowing, our system always asks if the user wants to explore a topic further or if they'd like an abstract generated. This gives users the choice to continue the discussion or move to abstract generation. Our system relies on three reliable source links from genuine scientific databases for accurate information. This ensures that the information it provides is credible and trustworthy.

We've also told our system to avoid repeating itself, quoting itself, or making things up, maintaining the integrity of the conversation. In terms of language, our system maintains a semi-formal tone. It uses complex terms from scholarly literature when appropriate, but it also aims for clarity to facilitate engaging conversations. Our system is designed to continuously learn and adapt. It takes cues from user feedback and new information to improve its responses. As an input variable for the *PromptTemplate* it uses *input history*. Worth mentioning that the history input was added as a safety net since we are already using *MemoryBuffer* function to keep track of 10 message exchanges between the system and the user. The experiments when we remove *history* from the prompt yield shallower and generic answers compared to when we include it in the prompt. Ultimately, the CRS builds on the user's questions to create meaningful discussions that stay relevant to the ongoing conversation. For example, if a user asks about quantum entanglement, our system would start a detailed conversation about the topic within the context of the ongoing discussion. This ensures that the information provided is not only scientifically accurate but also relevant to the user's queries. There was also a one-shot example (see Table 5.1) of a real abstract in order to showcase the formatting and semantic level expected in the generated abstract.

Table 5.1: Few Shot example used in the prompt

User: Generate a scientific abstract on {topic} **Assistant:** Here is a scientific abstract on {topic}: Title: "Impact of Climate Change on Coral Reefs" Context: "The research paper investigates the consequences of rising ocean temperatures and ocean acidification on coral reefs. The study combines field surveys, laboratory experiments, and computer modeling to assess the extent of coral bleaching, biodiversity loss, and ecosystem resilience. The findings highlight the urgent need for global efforts to mitigate climate change impacts on coral reef ecosystems." Generated Abstract: "The research paper examines the effects of climate change on coral reefs, focusing on rising ocean temperatures and ocean acidification. Through an interdisciplinary approach that integrates field surveys, laboratory experiments, and computer modeling, the study elucidates the severe consequences of these stressors on coral reef ecosystems. The research reveals a distressing trend of coral bleaching and subsequent biodiversity loss, contributing to the degradation of these vital marine ecosystems. Furthermore, the investigation underscores the importance of immediate and collaborative global action to curb climate change and safeguard the future of coral reefs.
Prompt: The conversation ends when the user says "bye" **User:** bye **Assistant:** Good luck in your research! **User:** Thanks, this is enough **Assistant:** Good luck in your research!

There is also an example of how the conversation is structured, clarifying the roles of the **Human** which provides the *input* and of the **Assistant** which has to reply.

Finally, our system is attentive to signals from users, like when they say goodbye, prompting an appropriate conclusion to the conversation. This ensures that the interaction remains user-centered and cohesive.

Table 5.2: Chain of Thought example used in the prompt

User input: "Please generate a 300-500 words abstract for a research paper titled 'Exploring the Role of Artificial Intelligence in Medical Diagnostics.' Consider the following points while crafting the abstract:"

- The increasing demand for accurate and efficient medical diagnoses.
- The integration of machine learning algorithms and deep learning techniques.
- Case studies showcasing successful AI-assisted diagnostic systems.
- The potential challenges and ethical considerations in AI-driven medical diagnostics.
- Implications for the future of healthcare and patient outcomes.

Generated Abstract: "The research paper delves into the transformative impact of artificial intelligence (AI) in revolutionizing medical diagnostics. With the escalating need for accurate and timely diagnoses, the study navigates the realm of machine learning algorithms and deep learning techniques to create sophisticated diagnostic systems. Through a comprehensive analysis of case studies, the research showcases remarkable achievements where AI-powered diagnostic tools have significantly improved accuracy and efficiency. Nevertheless, the journey is not devoid of challenges, as the integration of AI prompts ethical considerations and data privacy concerns. As the medical field continues to embrace AI, the paper contends that collaborative efforts between clinicians, researchers, and policymakers are pivotal in harnessing the full potential of AI-driven medical diagnostics. Ultimately, the research envisions a future where AI seamlessly augments medical professionals' capabilities, leading to enhanced patient care and medical outcomes."

Comparison of three prompt iterations

In Appendix A, three versions from different stages of the prompt refinement process. The first version found in Table A.1 is one of the first initial prompts. Some relevant mistakes worth mentioning, besides the grammatical errors, are the lack of adequate examples and overuse of adjectives such as "powerful" which heavily impact the prompt by not generating a proper abstract when asked and giving a general answer, as well as omitting referencing sources. Thus, the prompt design is heavily lacking regarding clear instructions and structure.

The second version found in Table A.2 shows that too much information confuses the model. The text was so long that I had to make two tables in the Appendix. When introducing Chain-of-Thought (see Table 5.2) and few-shot examples, the model began skipping the conversation and generating the abstract directly, sometimes even outputting one of the examples. However, the generated abstracts were more in-depth with a better structure (Methodology, Results etc.).

Last but not least, the final version from Table A.4 consisted of a moderate mix of detailing the output content, one shot examples, conversation structure, source retrieval, and specific enough abstracts. The trade-off was made so that the balance between the completion of tasks and the conversational aspect can be simultaneously achieved.

5.4 User Interface Development

This conversational recommender system's (CRS) user interface was created using Streamlit[1], an open-source, user-friendly framework. The user interface is designed to be simple and straightforward, facilitating a natural conversational experience.

Users are greeted with essential information about the user testing process, including the tasks they can perform and an overview of the user interface, upon landing on the homepage. This section introduces the CRS and explains to users what they can expect from their interactions with the system. The sidebar of the user interface contains two essential elements: the conversational recommender and an input field for the OpenAI key. Users can simply enter their OpenAI key to gain access to the system's features. The central element of the UI is the text input box, where users can type their questions or prompts to initiate conversations with the CRS. The system is programmed to introduce itself at the start of each conversation, establishing the stage for an engaging and user-friendly interaction. The user interface includes a hamburger menu on the left side for added convenience. This menu gives users the option to record the conversation, allowing them to review their interactions if they so choose.

The UI is deliberately designed to be simple and user-friendly, ensuring that users can engage in conversations with the CRS without difficulty. The combination of a clear introduction, an intuitive input field, and optional conversation recording enhances the overall user experience, making this platform accessible

5.5 Final prototype

Due to technical and knowledge constraints, a major part of the evaluation was changed. The final prototype included only hard prompts implementation. Thus it was split into two conditions, A and B, in order to gather insights from the users. The prototype constant features are the homepage, which consists the information on how to interact with the platform as well as the tasks; the OpenAI API key, which users can use their own or the one provided by the researchers; the chatbot pages, which each consist of an introductory text, a system message introducing "Iris" the chatbot and a text input field.

- **Condition A**: consists of two chatbots, *basic abstract generator* A.3 and *Chat with your PDFs*. Users could use one or both chatbots in order to gain insights into concepts from papers they are interested in and ask the chatbot to generate an abstract on a specific topic. The PDF chatbot also indicates the source of the answer, such as the paragraph from the uploaded PDF, acting as the source retrieval task. Condition A represented a simple, basic Q&A style that uses a minimum of conversation and offers quicker solution. We assume this condition will work best on researchers who already have a clear topic of research set. The basic chatbot uses half of the hard prompt, without the conversational parts (ending question, explaining concepts, etc.)

- **Condition B**: also two chatbots, however, the condition is conversational oriented. The *context aware chatbot* A.1 has all the main features we initially planned for in the beginning, such as conversational abilities, memory, source retrieval, and

[1]https://streamlit.io/

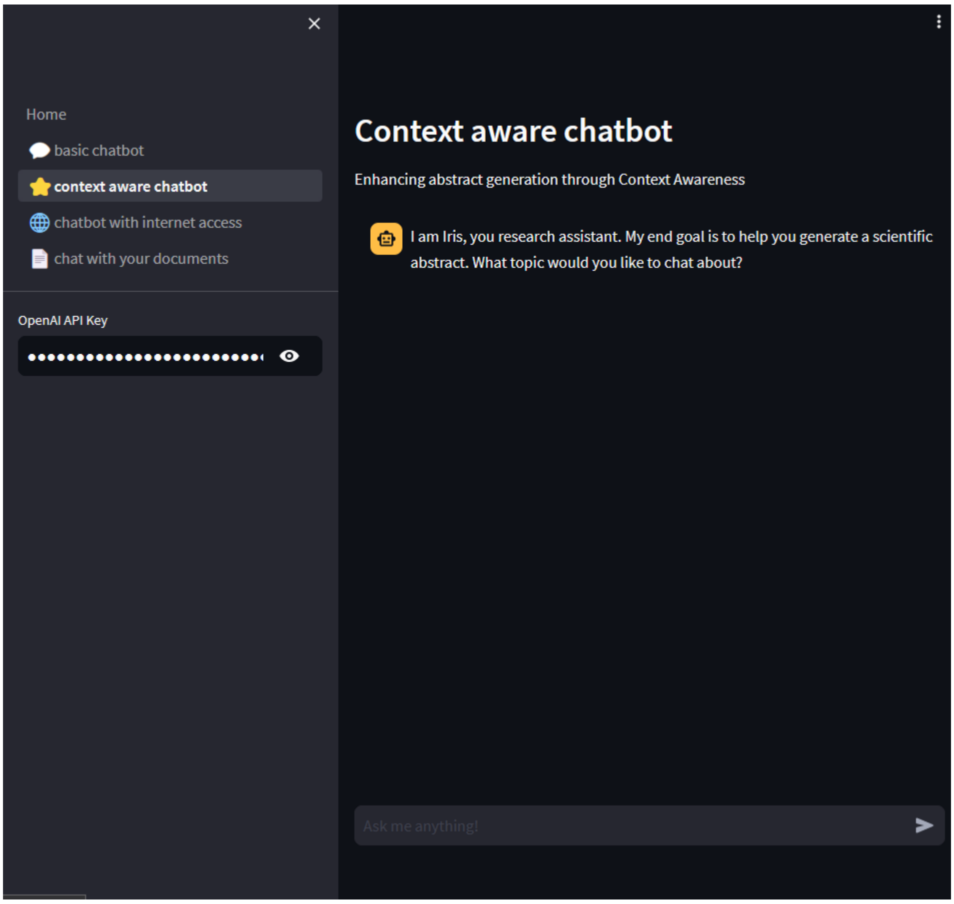

Figure 5.1: UI design

refined abstract generation based on conversation. We coupled this chatbot with a *Internet agent* A.2 which provides the latest information on any question. This was mostly done to tackle the issues of OpenAI API having access only to data up to 2022.

Therefore, we deviated from the original plan, where two chatbots should have been compared, the soft and hard prompted ones. The 4 hard prompted chatbots were selected in order to meet in the requirements from IRIS.AI, which considered having a PDF chatbot important for their product. Also, comparing a basic chatbot with just generates and abstract no matter the input with context-aware chatbot which can have a conversation, won't make it a fair comparison due to the absence of information. The context-aware chatbot can answer in detail any questions (except the new information dated after Nov 2021). Hence, the combination of having: **Condition A:** Q&A chatbot with "Basic chatbot" and "PDF chatbot" and the **Condition B:** Conversational chatbot with "Context Aware chatbot" and "Internet chatbot". Both Conditions are capable of generating abstracts similarly, however, A is less conversational and can answer any questions regarding specific PDFs uploaded by user, while B has the conversational element and can talk about anything up to Nov 2021. For more recent information, the internet chatbot is used similarly as the summarization tool part of Google for highlighting the primary information. Only the basic and the context-aware chatbot are prompted to generate the abstract, the other two are seen as adjunct tools.

Chapter 6

Evaluation

As mentioned in Implementation chapter 5, the soft prompts were not included in the user evaluation prototype, as originally planned, due to technical obstacles. For the explanations of the conditions, look up Section 5.5 which explains the choice of chatbots and their characteristics. There still remain two evaluation methodologies, user evaluation and benchmarking. The former gives us insights into the user experience with the CRS as well as outlines the actual needs of this specific target group (academia and researchers). The latter is a means of evaluating the semantic quality and accuracy of the generated abstract between soft and hard prompts. We also ask the experts to take a look at the generated abstract to evaluate if they are specific, sensible, and useful enough to be considered scientific abstracts. Table 6.1 shows how each objective is evaluated in this research.

Table 6.1: Evaluation methods

Objective	**Condition**	**Evaluation method**	**Participants**
User experience	**Condition A:** Q&A chatbot with "Basic chatbot" and "PDF chatbot"	User Evaluation SASSI	3
	Condition B: Conversational chatbot with "Context Aware chatbot" and "Internet chatbot"	User Evaluation SASSI	3
Content quality	**Hard Prompts**	SSA and benchmarking	3
	Soft Prompts	Benchmarking	-

6.1 Participants

6.1.1 Selection

SASSI: The user studies included only 6 participants which met the criteria mentioned in Chapter 3. Four are following a Master program and two are mid-way through their

PhD program. 2 of them attend the University of Twente and the others attend Aalto University. Initially, a higher number of participants was planned for the user evaluation, but 4 participants could not make it due to scheduling issues while 2 others had withdrawn their participation due to personal emergencies.

The recruitment of the participants was done via personal recruitment, where the participants who confirmed their participation were sent the information brochure and informed consent before starting the study.

Becuase of the low number of participants 3 participants experienced only one of each condition. In Chapter 5 all the changes to the initial plan are explained, as well as which conditions were used for user evaluations. In total, 6 participants were personally recruited via convenience sampling. No other demographics of the participants, besides which study they are following or which domain are they working in, were collected in this study to make sure that no identifiable information can be deducted about the participants. No gender quota or age range influences the participant's selection. The data collection took place over three days in an online setting, with the participant having the convenience of completing the study at any given time.

SSA: For the SSA, as mentioned in Chapter 3, we asked 3 experts to look at the generated abstracts from the SASSI user study, with the tasks of assessing the usefulness, specificity and sensibility of 4 of the system's answers (one answer for each chatbot). All 3 had to evaluate the same outputs. The experts were selected through convenience sampling and all 3 of them are industry experts in the field of Computer Science, with more than 10 years of experience and currently working at Iris.AI. however, they asked their personally identifiable information to not be published for this study. After the SASSI user evaluation was done, the experts were asked to fill in the SSA questionnaire.

6.1.2 User testing

To evaluate the prototype's effectiveness, the 6 participants were representative of the target user group. The testing phase provided valuable data on the system's performance and user satisfaction, to evaluate if the prompt tuning model met user expectations and requirements.

The participants received an online questionnaire and were asked via to complete the following two tasks: **1. Try out both chatbots** and **2. Obtain from the chatbot a scientific abstract regarding your topic of interest**. Below the two task the link to platform was presented. All the other instructions on how to use the platform, what are the goals of the chatbot and how to interact with them were present on the homepage of the application. After the conversation with the CRS was finished, the users had to return to the main questionnaire and complete the SASSI questionnaire.

There were a few limitations to take into consideration in this study. One of the cases could be that a participant suffers from dyslexia or the terms used were misunderstood due to language barriers. Another one was technical difficulties with the questionnaire. The only available solution was to have the researcher's information (phone number, telegram and email) presented on all pages of the questionnaire. There was no monetary compensation involved in this study, only a cordial appreciation message for the participant's contribution to the study. In the Appendix B a snippet of the questionnaire. Note that the next page button will only appear after 6 minutes.

As the focus is on the answers that the participants provide, the conversation between the Conversation Recommender System and the participant is recorded. The conversation log can be automatically recorded in a JSON file. After participation, the conversation logs are kept in the researcher's cloud storage affiliated with the university. All participants consented to have the samples from the conversation published in this paper. After one year, the conversations will be deleted.

6.2 Benchmarking

The utilization of benchmarking plays a crucial part in evaluating the performance of our system. The selection of evaluation criteria is based on a thorough analysis of recognized methodologies in the field, coupled with a thoughtful assessment of the characteristics and goals of our research.

In this evaluation, we will assess the effectiveness of soft prompts in generating text-to-text as well as the quality of the output from a semantic perspective. This approach involves the utilization of commonly acknowledged metrics such as **Rouge** [1], and **F1** [2]. Choice of metrics has been made based on their significance within the natural language processing field and their efficacy in assessing the quality of the generated text. The Rouge metric is a crucial quantitative indicator that assesses the degree of overlap and resemblance between the text generated by a model and the reference abstracts produced by humans. Its incorporation lies in its widespread utilization in scholarly works and its capacity to encompass many facets of textual excellence, such as precision and recall. F1 Score, often used in the context of classification tasks, is employed in this study to evaluate the appropriateness and precision of the generated text under prompt tuning technique. The usage of this approach is based on its ability to effectively balance precision and memory, being a good metric for evaluating the quality of generated responses.

6.3 SASSI questionnaire results

To highlight the differences between the two conditions, the most relevant findings from the SASSI questionnaire are presented through each of its six sections. The Likert scale of SASSI ranges from 1 (strongly disagree) to 5 (strongly agree).

System Response

For condition A, participants reported that the system is almost completely predictable (M = 1.80, SD = 0.20), but still dependable (M = 4.00, SD = 0) which would explain why the statement "*The system makes few errors*" scored the highest (M = 4.25, SD = 0.36). Most statements regarding accuracy, efficiency, and reliability scored close to neutral (M = 3.30, SD = 0.47).

Condition B score overall between 3 and 4 on average. Participants did not perceive the contextually aware chatbot as being more reliable or efficient in how it interacts, with one remarkable statement being "*The interaction with the system is efficient*" which scored the highest out of the three conditions (M = 4.15, SD = 0,81). The main factor might be the expectations of participants regarding the system, where the system was seen as unpredictable and not acting according to what the participants expected (M = 3.75, SD = 0.73).

[1]https://huggingface.co/spaces/evaluate-metric/rouge
[2]https://huggingface.co/spaces/evaluate-metric/f1

Likeability

Condition A scores are high on ease of use (M = 4.25, SD =1.36) and on clarity (M = 4.25, SD = 0.85). Overall, the participants' scores were leaning towards positive values, showing that the system was perceived as pleasant and friendly.

For Condition B, scores were neutral, with the exception of "*I was able to easily recover from errors*" where participants disagreed almost completely with the statement (M = 2.00, SD =1.24). This might be because of the memory feature where if multiple topics are discussed in a short span of time, the system has a harder time not relating the topics to one another in a specific domain context. Condition B scored positively on friendliness (M = 4.0, SD = 0.81) and pleasantness (M = 3.66, S D= 0.94).

Cognitive demand

Overall, the scores between the two conditions were similar, with the exception that Condition A was ranked the best by being the easiest to understand on how to use (M = 4.10, SD =0.35). The main difference between the two conditions is that while interacting with a conversational system requires more attention from the participants, it evokes more emotional responses from the participant (eg. calm, tense) compared to a question and answering system (Condition A). Still, this result is not outstanding to be considered significant yet due to the large variance in participants' answers. If the perception of the ECA and expectation of the conversation were pre and post-evaluated, then we could have had better insights into the emotional response which directly impacts the cognitive demand.

Annoyance

Even though Condition A was based on question and answering, with minimal conversation capabilities, it was perceived as being the opposite of boring, less repetitive than Condition B, and scored the highest on the system of use (M = 4.25, SD =0.36). System of use mainly encompasses usability and ease of use, thus participants found Condition A easier to interact with.

With an average score differential of 0.25 points less, Condition B performed similarly to the formal condition, with the only major exception being that it was perceived as more irritating (M = 3.30, SD =0.47) compared to Condition A (M = 2.15, SD = 1.25) based on the mean scores. The system's perception as repetitive and irritating is most likely due to the technical implementation of the elaboration system as well as condition B offering more information outlets, from access to the internet to the context-aware chatbot. Too many options might have overloaded the users leading to annoyance, as well as restarting the conversation when the chatbots are switched.

Habitability

Surprisingly, both conditions scored similarly in the habitability section, even for statements such as "*I sometimes wondered if I used the right word*" and "*I always knew what to write to the system*". Condition A proved to be the most consistent out of all two because participants could more easily keep track of the conversation (for the statement "*It is easy to lose track of where you are in an interaction with the system*", M = 2.13, SD= 0.43, compared to M = 3.50, SD= 0.69 condition B).

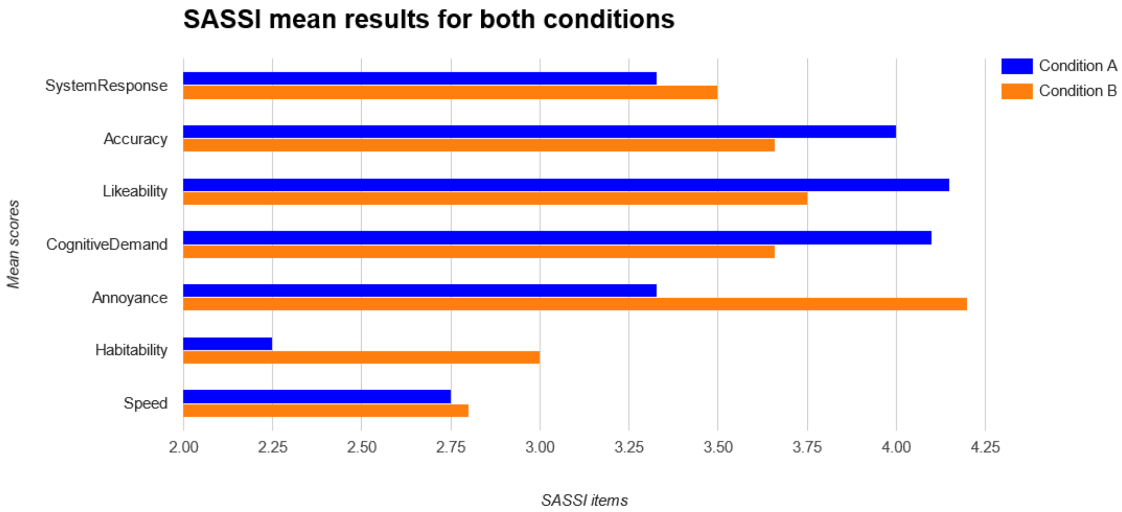

Figure 6.1: SASSI results

Speed

Participants were neutral about the systems' speed, leaning towards the systems having slower processing time of the answer (For "*The interaction with the system is fast*", all three conditions had an average of around 2.75).

Average completion time

It is worth mentioning the average completion time of each condition, which can be easily confronted in Table 6.4. For the condition A questionnaire, participants spent 15.5 minutes on average, with only 7.3 minutes on average using the chatbot. With Condition B, the average is 16.3 minutes; while 8.7 minutes is the average spent with the chatbot.

Condition	Average completion time	Chatbot time
A	15.4 m	7.3 m
B	16.3 m	8.7 m

Table 6.2: Average completion time of each of the two conditions.

6.4 Feedback

Most of the feedback concerning the Conversational Recommender System (CRS) primarily revolved around its content. For instance, respondents expressed that, "*The system often fails to capture the specific aspects of research mentioned during the conversation. For instance, it does not always assess the right methodology used in a paper and it does not interpret the results completely right, even though these aspects are highly relevant for the abstract*" and "*The generated text used os often too rigid and do not provide room for expressing personal opinions. An in-editing chatbot response would be great*". In terms of speed, one participant suggested, "*It should be quicker and more concise when processing* *multiple documents.*" Consequently, it can be deduced that the general format and the closed-ended questions represent notable weaknesses of the CRS.

Feedback pertaining to the Technical Implementation of the CRS, particularly the erasure of the conversation if the chatbots are switched in between, was a recurring theme. This frequently led to participant interruptions and limited their ability to formulate queries, especially when they started the whole conversation from the beginning. For instance, one participant commented, "*It's quite enjoyable. However, the interaction should allow me to edit my queries when I need to provide a more detailed response.*," or "*I wished it could have been faster in generating the answers, like instantly or write them word by word like chatGPT.*" Furthermore, participants noted the limited UI interaction from the system and the absence of features. One participant remarked, "*It seems like the system is designed for a more natural conversation, but it doesn't ask any specific questions for that context. I believe it could be enhanced by employing different questions and understanding of the language and possibly asking multiple follow-up questions within a given topic.*" This observation raises intriguing questions about the role of the CRS in text generation in specific domains and whether a natural conversation is warranted in such a scenario.

Issues raised by participants concerning Condition B mostly were linked to the elaboration source retrieval function and query processing of the user input. For example, participants indicated that "*The system appears to require very specific terms to continue the conversation. However, these terms may not always be easily recognizable or present in the user's response. When the system rejects an answer and requests more specific information, it should specify which terms are necessary for a better response,*" or "*It would be beneficial if the system can retrieve specific knowledge from the links provided. This way, when a user mentions something, the agent can understand the context even more in depth*".

Most participants voluntarily gave feedback to the researcher post evaluation, summarizing that while the generated abstracts are vague and shallow especially when discussing results, they are better formulated than their own abstract written by them for their research. We suggest for future studies to request the participants to write their own abstract and then compare it with the generated one to assess the quality of the content. The link prioritization was seen as a nice addition, however, the linked papers are the most commonly cited ones, and half of the participants did not gain significant new knowledge from them.

6.5 SSA

First, when examining each condition separately, it becomes clear that the ways in which the experts have assessed the system's responses vary. To clarify, we presented each of the experts with four interactions of the human system from all four chatbot conditions. The task of the experts was to rank the SSA of the system's response so in total they ranked four responses. The choice for Specificity, Sensibleness and Usefulness were either 1 or 0. All experts were presented with the same responses, thus there were cherry-picking biases involved by the researcher when selecting which interaction from the user reflects the best the ability of the chatbots. For the future, we recommend having two researchers who can pick the examples from the participants' conversations.

Looking at the results, the SSA does not offer us the expected insights into the quality of the system's responses. This could be because the experts did not interpret the questionnaire in the way that it was intended. As Adiwardana et al. (55) point out, the sensibility metric has been included as an extension to the specificity one, meaning that a sensible answer implies it is also specific. However, in the case of this research, there is a chance that the experts have understood the SSA questionnaire in a different way, making it problematic to state general conclusions on both the sensibility and specificity of the system responses.

Still, this research includes the usability metric, which proves to be a valuable addition. Except for the expert from the business department, there is a positive trend in usability looking at the averages from the two conditions. For instance, one experts thinks the response through Condition A are 0.45 useful and the responses through Condition B are 0.74 useful.

This positive trend is also noticeable in the total averages per evaluation condition, which can be seen in Table 6.3. The average usefulness of the Condition B is 0.68. This is an 142% increase with respect to the Condition A.

	Average per metric		
	Sensible	*Specific*	*Useful*
Condition A	0.48	0.44	0.27
Condition B	0.46	0.40	0.67

Table 6.3: Averages of the SSA metric per condition.

6.6 Soft Prompts evaluation

In this section, we present the evaluation results of our implementation, which focuses on assessing the effectiveness of soft prompts compared to hard prompts in the context of generating text-to-text responses for scientific paper recommendations. Only one task could be evaluated, the generation of an abstract. Our evaluation methodology is designed to provide insights into the performance of our approach and compare the model's Rouge-1 and F1. Furthermore, SciBERT was not a good choice of model for text generation, thus we excluded it from the results, comparing only GPT2 and Falcon. The RAFT dataset was used as the reference and the predictions were the 50 generated abstracted each by the models based on the title of the papers from RAFT dataset.

ROUGE

Our analysis using the Rouge metric revealed promising results. The metric computed the overlap and resemblance between the text generated by our soft prompt-based system and reference summaries produced by the authors of the articles. For Rouge summarization and text generation tasks, a score over 0.45 is considered good while under 0.35 is underperforming. We found that our system produced on average around 100 words more words than the initial abstract, with the appropriate linguistic terms and structure for a scientific abstract. One of the examples of the difference in the abstract generation can be seen in the Appendix A, between Table A.5 with the actual abstract and topic and Table A.6 which contains the abstract generated by Falcon model after prompt tuning. The hard prompts outperform the soft prompts in ROUGE, falcon-7b obtaining the best result. It is clear that prompting improves the model performance compared to no prompting.

Prompt	**Condition**	**Rouge-1**	**F1**
Soft	falcon-7b	0.36	0.38
	GPT2	0.28	0.21
Hard	falcon-7b	0.40	0.39
	GPT2	0.35	0.24
No Prompt	falcon-7b	0.33	030
	GPT2	0.27	0.22

Table 6.4: ROUGE and F1 scores

F1 score

While the F1 Score is typically used in classification tasks, we adapted it to evaluate the appropriateness and precision of the generated text under our prompt-tuning technique. Our results showed that the F1 Score provided valuable insights into the quality of the generated responses by the two models. The F1 Scores surprisingly had a significant gap in performance when looking at the soft prompts, with our Falcon implementation achieving an average F1 Score of 0.38. In comparison, GPT2 scored an average F1 Score of 0.21, suggesting a lower quality of generated responses. The F1 scores of both hard and soft for GPT2 were close to the no prompting scores. It might be mainly due to the large difference in the number of words between the generated abstract and the reference one or the GPT2 underperforming for text generation task.

Evaluation process conclusion

In summary, our implementation of soft prompts and hard prompts for text-to-text generation demonstrated some degree of success in improving the quality and efficiency of scientific abstract generation for the Iris.AI platform. The benchmarking metrics indicated that the choice of model for text generation is crucial in improving performance. The SASSI proved to be a good solution for assessing generative text AI, while SSA was maybe not the best evaluation method given the limited context.

While this evaluation has shown promising results, it is essential to acknowledge that further refinement and optimization may still be possible. Future work may involve fine-tuning the soft prompt generation process and exploring additional metrics to gain deeper insights into the system's performance. Also, creating a standardized questionnaire for evaluating the user experience as well as the quality of the system's answers. Lastly, a good idea would have been to implement an automatic way of verifying the links provided by the CRS for their authenticity instead of manually checking most of them when designing the final prototype. There were also a few design flaws that definitely impacted the evaluation, such as deleting the conversation when switching chatbot. These design mistakes happened due to changing the original methodology plan. Nevertheless, the positive outcomes of this evaluation lay a strong foundation for the potential adoption of prompting in enhancing user interactions and information retrieval in scientific research platforms.

Chapter 7

Discussion

There are three main takeaways from this research which we hope will be useful for future research.

- First, when interacting with the Conversational Recommender System (CRS), the outcomes of our study shed light on the particular preferences of our target user group, which is comprised of researchers following a Master or PhD. It is clear that the majority of these users choose to use the CRS as a quick abstract generator and Q&A chatbot, in particular for asking specific questions from specialized academic papers and generating abstracts. These users appear to have less faith in LLMs, which may be the root cause of their tendency toward a utilitarian, information-focused usage behavior. We assume in this research that the predisposition of the user community to depend on the CRS to swiftly condense information is symptomatic of a larger trend in the academic environment. This trend is one in which the efficacy and accuracy of AI-driven systems should meet precisely the requirements of academic and research-oriented work, thus conversational needs ranking at the bottom of priorities.

- Second, during the assessment of our CRS utilizing the SASSI and SSA questionnaires, a number of complex problems regarding the quality and specificity of the information that was generated were identified. These problems highlighted the need for a more thorough analysis of the user population and their specific requirements in order to tailor the recommendation system effectively. While we created a wishlist requirement for our ideal CRS, this set of items should serve as a starting point for future design of similar studies. We noticed that the users expressed their concerns, stating that they believed that the abstracts created by the CRS and the abstracts generated through prompt engineering did not live up to the ideal standard that is normally associated with their human-written equivalents. In this research, generated abstracts should be vague, since they are used as a query in the Iris.AI platform. Besides, the input is minimal so the PTM has to fill in the gaps with their own knowledge following the scientific abstract structure from the prompt instructions. However, it is relevant to manage user expectations and we hope future research can focus more on a thorough comparison in the content quality of the abstracts generated by soft and hard prompts.

- Compared to the difference between the two conditions in the user evaluation, there was not a big disparity in the benchmarking results between the hard and soft prompts, especially in the performance of falcon-7 b evaluated through ROUGE score. This disparity in perceived quality draws attention to the domain adaptation in which the performance of the LLMs can be improved for text-generative tasks as well as the necessity of standardized user evaluation for such tasks. While we adapted SASSI and SSA for our research, there low number of participants skewed the results while still offering some useful insights into the UX of the CRS. Thus, we urge the adoption of user testing for conversational systems that rely on LLM and other NLP techniques in order to gain a deeper understanding of the actual performance of the system, not only drawing conclusions from numeric metrics.

7.1 Discussion of the results

Expanding on the two tasks of the CRS, it is clear that improving the system such that it can generate scientific abstracts and scientific sources as well as engage in discussion offers a complex set of challenges. Our prompt-tuned model performed well in the first element by creating coherent scientific abstracts. The structure of an abstract was present. However, the content was shallow, the abbreviations and basic concepts were over-described for the purpose of an abstract, and the interpretation of the results was most of the time absent or theoretical instead of statistical and concrete.

The CRS had significant challenges in providing a conversational experience that was agreeable to the user. From the feedback on Condition B, the conversational chatbot, the users find that the explicit and long answers contribute to losing track of the conversation, which increases annoyance and decreases ease of use. This subtle remark highlights the possible trade-off that is inherent in design decisions, highlighting the necessity of striking a careful balance between these two aspects of the design. The major use case of the system as well as the user's expectations, emerge as important factors that require careful design decisions. Unfortunately, the erasing conversation feature (see Chapter 5) when switching between the two chatbots was a poor design choice, even though the motivation behind was increasing the speed of the models and not overloading the memory conversational chains. The emphasis that is placed on conversational functionality versus quality of the generated abstract in a comprehensive recommendation system should be carefully weighed, depending on the target group that the system is designed for and the requirements that they have in particular.

Improving user experience

Incorporating user feedback and preferences into the development process can help refine the system's functionality and ensure a more satisfactory user experience. The way the prompt refinement was executed was subjective and included only the opinion of the researchers on what it is the most adequate, perfect prompt for this goal. User center design in LLMs design should be demanded more, since benchmarking models it is not enough to justify the quality of the outputs, especially in generative tasks for specific domains.

In addition, the results of the SSA questionnaire highlighted the significance of including both specifics and generalizations in the information that was provided. Users have shown a need for information that not only corresponds to their research interests but also has the distinctive features of clarity and accuracy. The repercussions of these discoveries are not limited to simple user choices; rather, they highlight the importance of performing careful user research and understanding user needs in order to design effective information systems. This not only enhances the overall user experience but also increases the system's value and relevance in supporting research endeavors. Therefore, it is crucial for designers to continuously evaluate and improve the system based on user feedback and evolving research needs.

Besides, we used SASSI and SSA which worked relatively well for the purpose of this research evaluation. However, we can not guarantee this evaluation style is scalable and future-proof for further generative LLMs user evaluations.

LLMs and methodology modifications

Mainly, prompts should guide the models in such a way that they can recreate the few-shot examples (for hard prompts) and the training data output (soft prompts) with increased precision, such as introducing statistical or numeric results, interpreting the choice of methodology and not over-explaining basic concepts. Understandably, such results prove to be difficult to achieve if the user input is solely the title of the research without any conversation. As seen in the prompt-engineered CRS experiment, relying on a framework such as LangChain to take into consideration the chat history can create a black box effect where the generated abstract becomes too long and rips us of any certainty that the abstract took into consideration the research specifics from the user input. For soft prompting, training the model on a smaller dataset which takes the title of the paper as input, has consequences such as limited knowledge and problems in adapting to other tasks (such as Q&A) when the main tasks were mainly generative text.

Regarding the choice of LLMs, we acknowledge it was not the most coherent implementation of the proposed methodology, which had a negative impact on the results. While in general, generative models such as Falcon and GPT3 performed well, the soft prompting evaluation could have been better by including the conversational and source retrieval tasks, not only abstract generation. The evaluation of the LLMs was not clear and susceptible to subjective bias. Thus, we stress the growing need for a standard way of benchmarking both hard and soft prompts in further studies.

7.2 Future research

The findings of this study have some significant implications for the course of future research and development in the field of conversational recommender systems and prompting techniques, in particular when adapted to the requirements of users in academic and research settings. In the first place, our findings emphasize the importance of the requirement of a nuanced strategy that takes into consideration the various usage circumstances and expectations of the users. In the course of their work, researchers and academics frequently navigate the complex environment of information retrieval, which might include activities ranging from a cursory fact check to an in-depth investigation. Future research attempts should aim to incorporate a spectrum of features inside the CRS, such as conversing with the papers, generating human-like abstracts and link prioritization of more specific articles. This will allow the system to accommodate both the speedy creation of abstracts that users are looking for, as well as the nuanced conversational help that users may want while discussing complex subjects. However, we highlight that the design of the CRS should be minimalistic, including the users or experts in the design process and streamlined as much as possible for reproducible purposes.

In addition, the comparative examination of soft and hard prompts offers a nuanced viewpoint on the design choices that are inherent in conversational recommender systems. According to the findings of our research, prompt-tuned models perform very well in certain activities while being constrained in others. Because of this duality, it is necessary to conduct an all-encompassing investigation of hybrid methods, which may capitalize on the advantages presented by both prompt tuning and prompt engineering. In further study, a greater focus should be placed on the optimization of prompt methods, with particular attention paid to the dynamic interplay between prompt formatting, LM fine-tuning, and user interactions. Prompt engineering should be transparently assessed according to an agreed standard, testing what kind of instruction creates the difference in the model's behavior. Prompt tuning is rapidly become an outdated technique, with other techniques, such as adapters and P-tuning being more efficient and easier to optimize. Thus the choice of which technique should be implemented in testing soft prompts should be properly justified (not like in our research). In addition, research can investigate techniques for dynamically altering prompt tactics based on user behavior and the context of the situation, such as having both hard and soft prompts in the CRS, which would increase the flexibility of the system and a cleared approach to evaluating the models.

As a note, the topic of whether or not users can put their faith in Language Models has emerged as a critical component that calls for in-depth research. It is possible to create more effective conversational recommender systems by first gaining an understanding of the underlying elements that lead to user trust and how it affects user behavior in real-world circumstances. In the future, research should investigate several ways for establishing and sustaining trust, not only relying on the gathered requirements for developing the system. Some of these tactics might include transparent system behavior, open source models, fewer external dependencies, strong privacy safeguards, and tools for user control over model outputs. In addition, the study should investigate the role of explicability in gaining user trust. It is possible to increase the users' level of trust in the system's proposals by giving them explanations that are easy to grasp for the recommendations given by language models instead of letting user expectations influenced by societal norms impact the perception of the generated results. The investigation of the influence of user input and the incorporation of that feedback into the learning process of the model can also be extremely important for both the establishment and maintenance of trust.

7.3 Ethical implications and limitations

It is of the utmost importance that the limitations that are inherent to our study be recognized and addressed. One of our most significant limitations is the technical abilities and lack of knowledge of the authors which led to a bad implementation of the prompting techniques. Furthermore, the specialization of our user group, is predominately made up of persons from the research and academic communities. It's possible that members of this demographic have their own particular tastes and standards, which would make it difficult to extrapolate our findings to cover a wider range of user communities. The user cohort for future study should be diversified to embrace a wider range of backgrounds and requirements in order to provide a more thorough knowledge of user preferences. This should be a goal of future research in user evaluations of LLMs.

In addition, ethical issues must not be ignored throughout the process of developing and deploying CRS. When using LLMs in a responsible manner, one must be vigilant in minimizing biases and ensuring that suggestions do not perpetuate unjust inequities or disinformation. This requires one to pay close attention to details from eliminating subjective bias in prompt design to adequately fact-checking the LLMs output by following a transparent pipeline for the generated answers. In order to address these issues, it is imperative that future research places a priority on the creation of ethical principles as well as procedures for continual monitoring and improvement of prompting techniques implementations and user evaluations. In addition, it is essential to include a wide variety of stakeholders, such as users and specialists hailing from a variety of fields, in the process of designing and evaluating CRS. This method of working together can assist in identifying and handling possible ethical concerns from a variety of viewpoints, which will ensure that the system respects the user's autonomy and encourages inclusion. In addition, measures of transparency and accountability should be put into place in order to encourage confidence and make it possible for users to have a clear knowledge of how the data they provide is being utilized in the process of making recommendations, eliminating the preconceived expectations of the interaction with the CRS.

Chapter 8

Conclusions

In conclusion, this master's thesis undertook an inquiry into the efficacy of soft and hard prompt techniques in the scientific field, particularly in relation to the development of conversational abstracts. This study provides a overview of conversational recommender systems, a field that is crucial for academics and researchers seeking efficient and personalized information discovery tools. It tackled some of the important challenges of assisting users in writing research abstracts by using a unique approach that integrates prompt engineering and prompt tuning into a CRS. The research utilized two evaluation methodologies, which involved the integration of user research and objective performance criteria. This approach was employed after examining the requirements, advantages and drawbacks of both prompt categories.

This inquiry yielded three significant contributions. Initially, a comprehensive compilation of necessary conditions and hypothetical scenarios was constructed based on interviews with IRIS.AI employees which offered some insights into the technical and use outlines for this specific research combined with a framework for prospective investigations from the literature review. This wishlist spans a range of technological, user-centric, and functional viewpoints, hence serving as a starting point for future research, be it in the industry or academia.

Furthermore, an investigation into user studies provides insights into the tastes and habits of the target population, which consists of academics and researchers. The results highlighted a preference for interactions that focus on factual information and question-and-answer exchanges, rather than extensive and conversational encounters. One important takeaway was that matching prompt techniques with user preferences and domain-specific demands is necessary when considering what is the user's image of desired output.

Also, the research investigated the clarity and coherence of the produced abstracts, utilizing well-established metrics such as Rouge and F1 scores. Generating broader and vague scientific abstracts was expected given that we only inject a prompt instruction and user input, letting the model fill in the blanks with its knowledge. It was also the perfect output for our research application, since broader abstracts serve as better queries for a literature database such Iris.AI, mainly because it gives information about the context rather than specific terms which can be misinterpreted. The study uncovered a discrepancy between the outputs generated by the system and the expectations of the users. The aforementioned disparity shed light on the intricacies and benefits linked to the integration of prompting approaches inside a CRS.

Ethical issues and user trust also proved to be vital aspects of conversational recommender systems. To secure the appropriate deployment of AI-driven systems, ethical principles and procedures must be firmly defined, including the mitigation of prejudice and safeguarding of privacy. The trust serves as the foundation upon which user engagement and acceptance are constructed. Ensuring transparency in the decision-making process of CRS is crucial to address concerns related to the source of the generated answer and the thinking pattern of the LLM.

This study provides an important contribution to the field of conversational recommender systems, highlighting the importance of contextual comprehension and the necessity for flexible prompt tactics. One of the main conclusions drawn from the study is the significant realization that strategies for prompts need to be customized in order to correspond with user preferences and the specific requirements of the scientific field. Since LLMs are becoming a focus for future upgrades, as the academic and research sector demands accuracy and clarity in knowledge extraction, this research provides a list of follow-up improvements from the methodology down to evaluations. Consequently, the research lays the groundwork for the development of future systems that prioritize the needs and preferences of users, resulting in enhanced effectiveness.

Bibliography

[1] Building a Better Search Engine for Semantic Scholar | by Sergey Feldman | AI2 Blog.

[2] Adikari, S., Mcdonald, C., and Campbell, J. Little Design Up-Front: A Design Science Approach to Integrating Usability into Agile Requirements Engineering. 549–558.

[3] Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., Riedel, C. J., Hine, E., Ashurst, C., Sedille, P., Carlier, A., Noetel, M., and StuhlmÃ¼ller, A. RAFT: A Real-World Few-Shot Text Classification Benchmark. *arXiv* (Jan. 2022). arXiv:2109.14076 [cs].

[4] Anastas, J. W. Quality in Qualitative Evaluation: Issues and Possible Answers. *Research on Social Work Practice 14*, 1 (Jan. 2004), 57–65. journal: SAGE Publications Inc.

[5] Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019).

[6] Ben-David, E., Oved, N., and Reichart, R. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *arXiv* (Jan. 2022). arXiv:2102.12206 [cs].

[7] Bhandari, M., Gour, P. N., Ashfaq, A., Liu, P., and Neubig, G. Re-evaluating evaluation in text summarization.

[8] Biggs, M. The role of the artefact in art and design research. *International Journal of Design Science and Technology 10* (Jan. 2002).

[9] Bordes, A., Weston, J., and Usunier, N. Open Question Answering with Weakly Supervised Embedding Models. *Springer* (2014), 165–180.

[10] Bouraoui, Z., Camacho-Collados, J., and Schockaert, S. Inducing Relational Knowledge from BERT. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 05 (Apr. 2020), 7456–7463. Number: 05.

[11] Brown, and al. Language models are few-shot learners. *Curran Associates, Inc. 33* (2020), 1877–1901.

[12] Cai, W., Jin, Y., and Chen, L. Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems. *ACM* (Apr. 2022), 1–14.

[13] Cai, X., Liu, S., Yang, L., Lu, Y., Zhao, J., Shen, D., and Liu, T. COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers. *Journal of Biomedical Informatics 127* (Mar. 2022), 103999.

[14] Chen, Q., Du, J., Allot, A., and Lu, Z. Litmc-bert: Transformer-based multi-label classification of biomedical literature with an application on covid-19 literature curation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 19*, 5 (Sep 2022), 2584â2595.

[15] Chen, Y., Liu, Y., Dong, L., Wang, S., Zhu, C., Zeng, M., and Zhang, Y. AdaPrompt: Adaptive Model Training for Prompt-based NLP. *Association for Computational Linguistics* (Dec. 2022), 6057–6068.

[16] Chuang, Y.-N., Tang, R., Jiang, X., and Hu, X. SPeC: A Soft Prompt-Based Calibration on Mitigating Performance Variability in Clinical Notes Summarization. *arXiv* (Mar. 2023). arXiv:2303.13035 [cs].

[17] Deng, Y., Zhang, W., Xu, W., Lei, W., Chua, T.-S., and Lam, W. A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems. *arXiv* (Apr. 2022). arXiv:2204.06923 [cs].

[18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Association for Computational Linguistics* (June 2019), 4171–4186.

[19] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *arXiv* (Feb. 2019). arXiv:1811.01241 [cs].

[20] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *arXiv* (Feb. 2019). arXiv:1811.01241 [cs].

[21] Erkan, G., and Radev, D. R. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research 22* (Dec. 2004), 457–479.

[22] Ermakova, L., Bordignon, F., Turenne, N., and Noel, M. Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences. *Frontiers in Research Metrics and Analytics 3* (2018).

[23] Feldman, S. Natural Language Processing in Information Retrieval.

[24] Floridi, L., and Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines 30*, 4 (Dec. 2020), 681–694.

[25] Gao, C., Lei, W., He, X., de Rijke, M., and Chua, T.-S. Advances and challenges in conversational recommender systems: A survey. *AI Open 2* (Jan. 2021), 100–126.

[26] Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners.

[27] Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., and Zhang, J. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. arXiv:2303.14524 [cs].

[28] Glaser, B., and Holton, J. Remodeling Grounded Theory. *Forum: Qualitative Social Research 32* (Jan. 2007), 47–68.

[29] Gupta, A., Basu, D., Ghantasala, R., Qiu, S., and Gadiraju, U. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. *ACM* (Apr. 2022), 3531–3540.

[30] Gupta, S., and Gupta, S. K. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications 121* (May 2019), 49–65.

[31] Hao, S., Tan, B., Tang, K., Ni, B., Shao, X., Zhang, H., Xing, E., and Hu, Z. Bertnet: Harvesting knowledge graphs with arbitrary relations from pretrained language models.

[32] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-Efficient Transfer Learning for NLP. *arXiv* (June 2019). arXiv:1902.00751 [cs, stat].

[33] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* (Oct. 2021). arXiv:2106.09685 [cs].

[34] Iovine, A., Narducci, F., and Semeraro, G. Conversational Recommender Systems and natural language:. *Decision Support Systems 131* (Apr. 2020), 113250.

[35] Jiang, T., Jiao, J., Huang, S., Zhang, Z., Wang, D., Zhuang, F., Wei, F., Huang, H., Deng, D., and Zhang, Q. PromptBERT: Improving BERT Sentence Embeddings with Prompts. *Association for Computational Linguistics* (Dec. 2022), 8826–8837.

[36] Johannesson, P., and Perjons, E. *An Introduction to Design Science*. Cham, 2014.

[37] Kassner, N., Dufter, P., and SchÃ¼tze, H. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. *Association for Computational Linguistics* (Apr. 2021), 3250–3258.

[38] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. CTRL: A conditional transformer language model for controllable generation.

[39] Kocabalil, A. B., Laranjo, L., and Coiera, E. *Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires*. BCS Learning 38; Development, Jul 2018.

[40] Kryscinski, W., McCann, B., Xiong, C., and Socher, R. Evaluating the factual consistency of abstractive text summarization.

[41] Kulak, D., and Guiney, E. *Use Cases: Requirements in Context*. Mar. 2012.

[42] Lester, B., Al-Rfou, R., and Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. *arXiv* (Sept. 2021). arXiv:2104.08691 [cs].

[43] Lester, B., Al-Rfou, R., and Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. *arXiv* (Sept. 2021). arXiv:2104.08691 [cs].

[44] Li, S., Zhu, Y., Xie, R., Tang, Z., Zhang, Z., Zhuang, F., He, Q., and Xiong, H. Customized conversational recommender systems. *arXiv.org* (Jun 2022).

[45] Li, X. L., and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation.

[46] Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries.

[47] Liu, J., Cohen, S. B., and Lapata, M. Discourse Representation Structure Parsing. *Association for Computational Linguistics* (July 2018), 429–439.

[48] Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Bras, R. L., Choi, Y., and Hajishirzi, H. Generated Knowledge Prompting for Commonsense Reasoning. *arXiv* (Sept. 2022). arXiv:2110.08387 [cs].

[49] Liu, P., Zhang, L., and Gulla, J. A. Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems. *arXiv* (Mar. 2023). arXiv:2302.03735 [cs].

[50] Liu, P., Zhang, L., and Gulla, J. A. Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems. *arXiv* (Mar. 2023). arXiv:2302.03735 [cs].

[51] Liu, X., Gao, Y., Bai, Y., Li, J., Hu, Y., Huang, H., and Chen, B. PSP: Pre-trained Soft Prompts for Few-Shot Abstractive Summarization. *arXiv* (Oct. 2022). arXiv:2204.04413 [cs].

[52] Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., and Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. *Association for Computational Linguistics* (May 2022), 61–68.

[53] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv* (May 2023). arXiv:2303.16634 [cs].

[54] Logan IV, R. L., BalaÅ¾eviÄ, I., Wallace, E., Petroni, F., Singh, S., and Riedel, S. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. *arXiv* (July 2021). arXiv:2106.13353 [cs].

[55] Lu, H., Bao, S., He, H., Wang, F., Wu, H., and Wang, H. Towards boosting the open-domain chatbot with human feedback. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023).

[56] Lund, B. D., and Wang, T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries?

[57] Luu, K., Wu, X., Koncel-Kedziorski, R., Lo, K., Cachola, I., and Smith, N. A. Explaining Relationships Between Scientific Documents. *arXiv* (Aug. 2021). arXiv:2002.00317 [cs].

[58] MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., and Goharian, N. Simplified Data Wrangling with ir_datasets. *ACM* (July 2021), 2429–2436.

[59] Manakul, P., Liusie, A., and Gales, M. J. F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *arXiv* (May 2023). arXiv:2303.08896 [cs].

[60] Manzoor, A., and Jannach, D. Towards retrieval-based conversational recommendation. *Information Systems 109* (Mar. 2022), 102083. journal: Pergamon.

[61] Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., OâConnor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., and Moore, J. H. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining 16*, 1 (July 2023), 20.

[62] Mousavi, R., Raghu, T., and Frey, K. Harnessing Artificial Intelligence to Improve the Quality of Answers in Online Question-answering Health Forums. *Journal of Management Information Systems 37*, 4 (Oct. 2020), 1073–1098. journal: Routledge _eprint: https://doi.org/10.1080/07421222.2020.1831775.

[63] Narducci, F., De Gemmis, M., Lops, P., and Semeraro, G. Improving the User Experience with a Conversational Recommender System. In *AI*IA 2018 â Advances in Artificial Intelligence*, C. Ghidini, B. Magnini, A. Passerini, and P. Traverso, Eds., vol. 11298. Cham, 2018, pp. 528–538. Series Title: Lecture Notes in Computer Science.

[64] Nayak, A., Timmapathini, H., Ponnalagu, K., and Gopalan Venkoparao, V. Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words.

[65] OpenAI. ChatGPT (Aug 23 version). Large language model.

[66] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *arXiv* (Mar. 2022). arXiv:2203.02155 [cs].

[67] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. *Association for Computational Linguistics* (2001), 311.

[68] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems 24*, 3 (Dec. 2007), 45–77. journal: Routledge _eprint: https://doi.org/10.2753/MIS0742-1222240302.

[69] Peng, Y., Yan, S., and Lu, Z. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

[70] Petroni, F., RocktÃ¤schel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language Models as Knowledge Bases? *Association for Computational Linguistics* (Nov. 2019), 2463–2473.

[71] Pramod, D., and Bafna, P. Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review. *Expert Systems with Applications 203* (Oct. 2022), 117539.

[72] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving Language Understanding by Generative Pre-Training.

[73] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* (July 2020). arXiv:1910.10683 [cs, stat].

[74] Raghu, M., and Schmidt, E. A survey of deep learning for scientific discovery. *arXiv.org* (Mar 2020).

[75] Ricci, F., Rokach, L., and Shapira, B. Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA, 2011, pp. 1–35.

[76] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J. Recipes for building an open-domain chatbot. *arXiv* (Apr. 2020). arXiv:2004.13637 [cs].

[77] Rush, A. M., Chopra, S., and Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv* (Sept. 2015). arXiv:1509.00685 [cs].

[78] Ryan. Project Zero: A walk through Project Zero metrics. *Project Zero* (Feb. 2022).

[79] Sakor, A., Singh, K., Patel, A., and Vidal, M.-E. Falcon 2.0: An Entity and Relation Linking Tool over Wikidata. *Association for Computing Machinery* (Oct. 2020), 3141–3148.

[80] Scao, T. L., and Rush, A. M. How Many Data Points is a Prompt Worth? *arXiv* (Apr. 2021). arXiv:2103.08493 [cs].

[81] Schick, T., and SchÃ¼tze, H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. *Association for Computational Linguistics* (Apr. 2021), 255–269.

[82] Senadeera, D., and Ive, J. *Controlled Text Generation using T5 based Encoder-Decoder Soft Prompt Tuning and Analysis of the Utility of Generated Text in AI.* Dec. 2022.

[83] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *arXiv* (Nov. 2020). arXiv:2010.15980 [cs].

[84] Simon, H. A. *The sciences of the artificial.* Karl Taylor Compton lectures. Cambridge, 1969. OCLC: 4087.

[85] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to Sequence Learning with Neural Networks.

[86] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., and Le, Q. LaMDA: Language Models for Dialog Applications. *arXiv* (Feb. 2022). arXiv:2201.08239 [cs].

[87] Vargas, S., and Castells, P. Rank and relevance in novelty and diversity metrics for recommender systems. *Association for Computing Machinery* (Oct. 2011), 109–116.

[88] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv.org* (Jun 2017).

[89] Vu, T., Lester, B., Constant, N., Al-Rfouâ, R., and Cer, D. SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer. *Association for Computational Linguistics* (2022), 5039–5059.

[90] Walsh, I. Using quantitative data in mixed-design grounded theory studies: an enhanced path to formal grounded theory in information systems. *European Journal of Information Systems 24*, 5 (Sept. 2015), 531–557.

[91] Wang, Q., Downey, D., Ji, H., and Hope, T. Learning to Generate Novel Scientific Directions with Contextualized Literature-based Discovery. *arXiv* (May 2023). arXiv:2305.14259 [cs].

[92] Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., and Pfister, T. Learning to prompt for continual learning.

[93] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* (Jan. 2023). arXiv:2201.11903 [cs].

[94] Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., and Goldstein, T. Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery. *arXiv* (June 2023). arXiv:2302.03668 [cs].

[95] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv* (Feb. 2023). arXiv:2302.11382 [cs].

[96] Wu, N., Gong, M., Shou, L., Liang, S., and Jiang, D. Large Language Models are Diverse Role-Players for Summarization Evaluation. *arXiv* (Mar. 2023). arXiv:2303.15078 [cs].

[97] Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. Why Johnny Canât Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *ACM* (Apr. 2023), 1–21.

[98] Zangerle, E., and Bauer, C. Evaluating Recommender Systems: Survey and Framework. *ACM Computing Surveys 55*, 8 (Dec. 2022), 170:1–170:38.

[99] Zhang, Y., Chen, X., Ai, Q., Yang, L., and Croft, W. B. Towards Conversational Search and Recommendation: System Ask, User Respond. *ACM* (Oct. 2018), 177–186.

[100] Zhang, Y., Fei, H., Li, D., and Li, P. PromptGen: Automatically Generate Prompts using Generative Models. *Association for Computational Linguistics* (July 2022), 30–37.

[101] Zhao, T. Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *arXiv* (June 2021). arXiv:2102.09690 [cs].

[102] Zheng, C., and Huang, M. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv.org* (Sep 2021).

[103] Zhong, Z., Friedman, D., and Chen, D. Factual Probing Is [MASK]: Learning vs. Learning to Recall. *arXiv* (Dec. 2021). arXiv:2104.05240 [cs].

Appendix A

First appendix

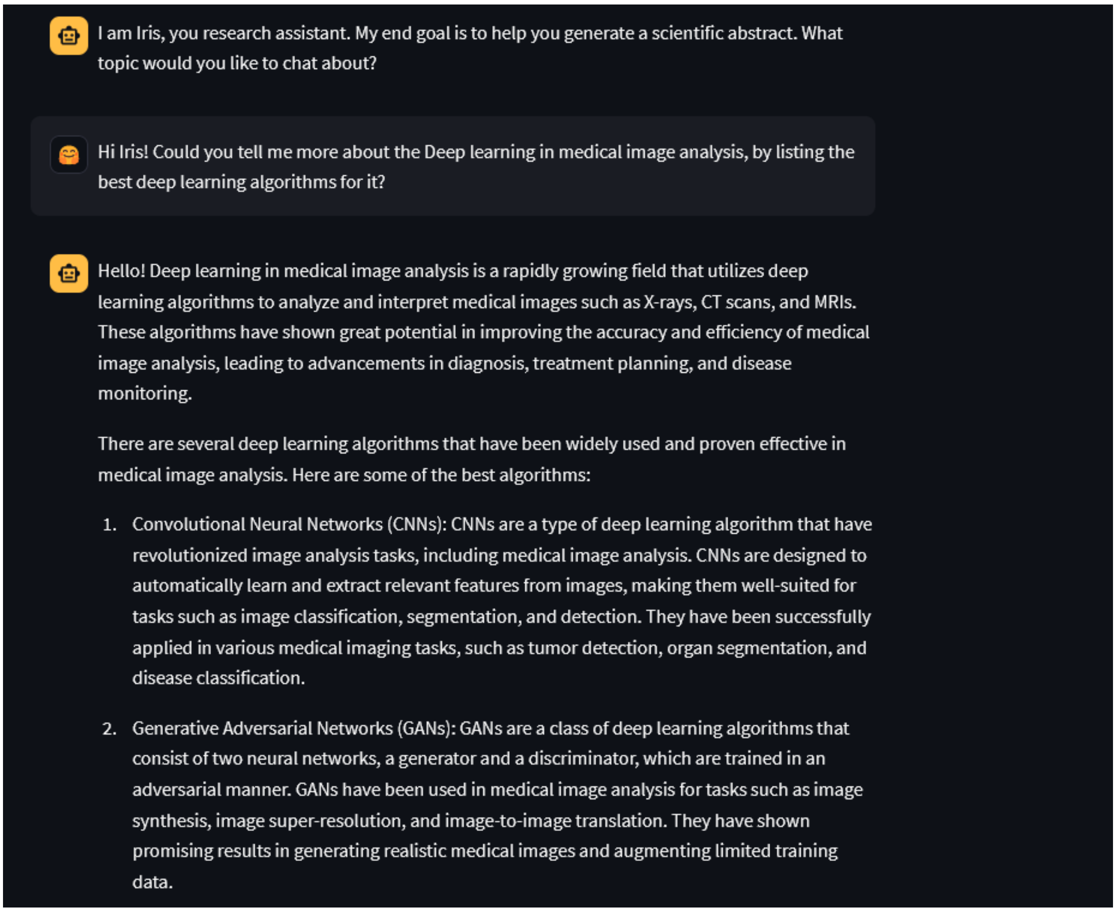

Figure A.1: Context chatbot

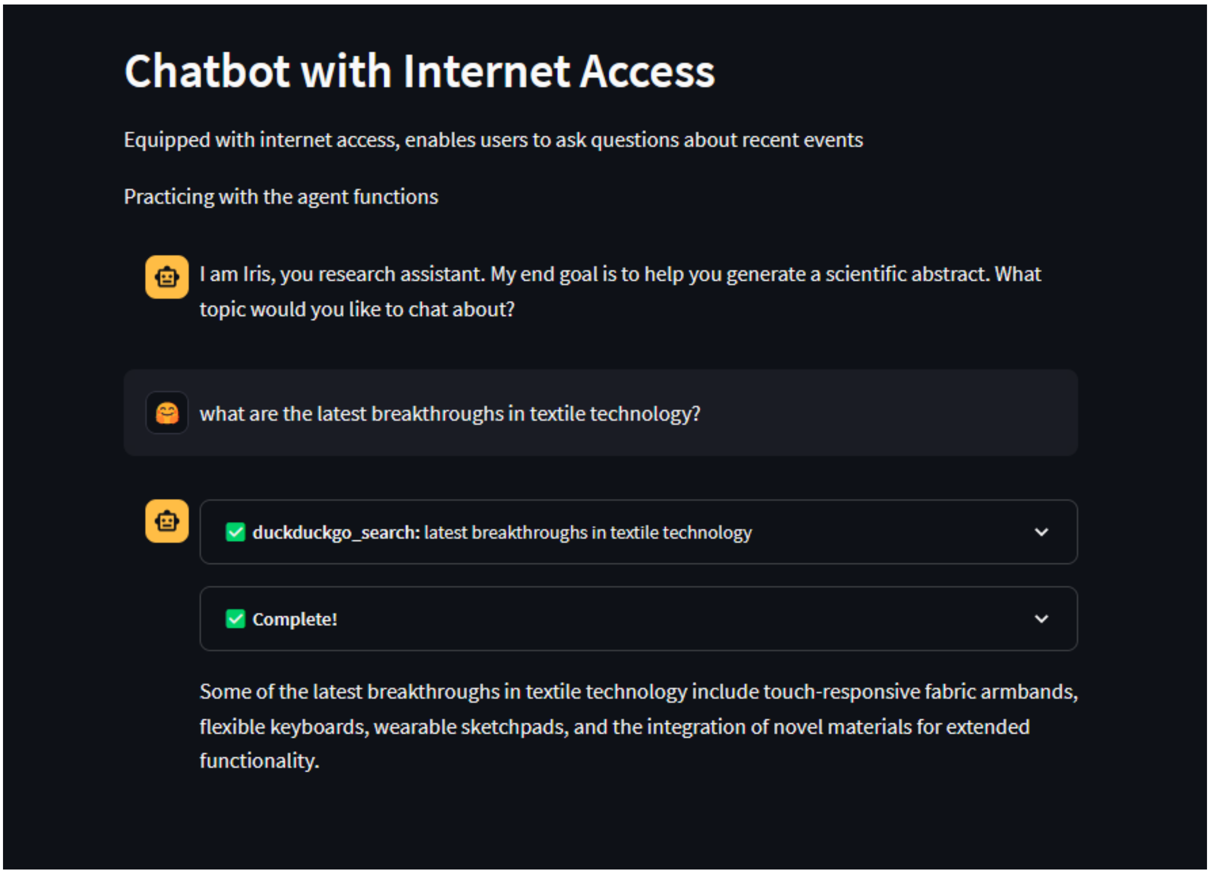

Figure A.2: Internet chatbot

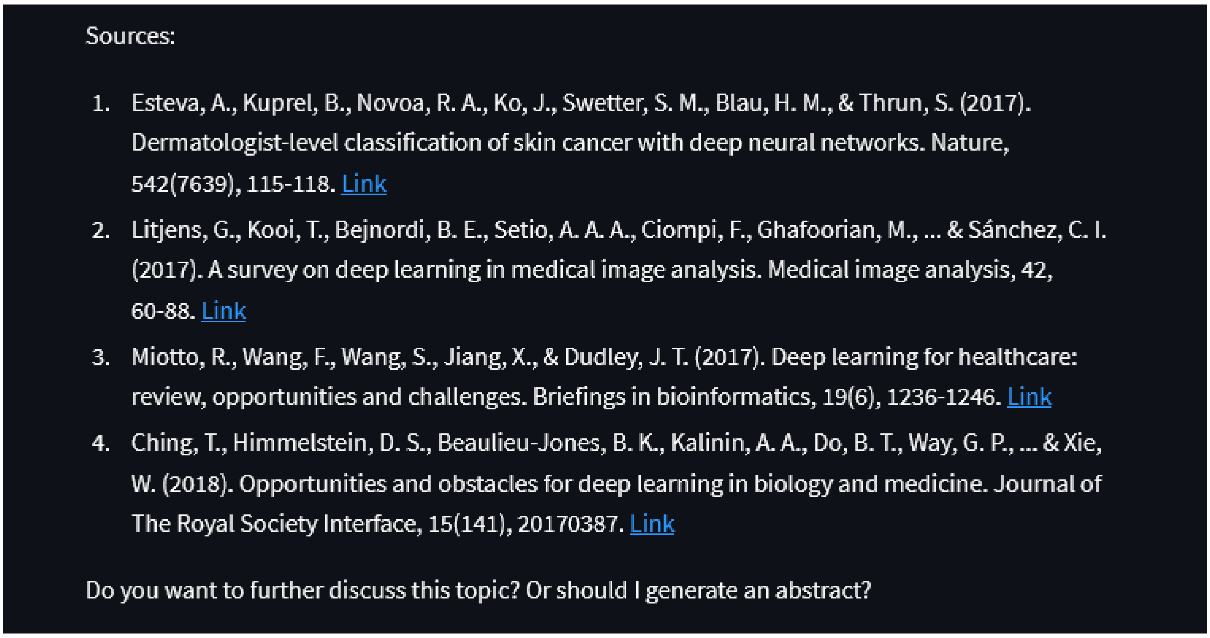

Figure A.3: Basic chatbot source retrieval

Table A.1: One of the first iterations of the prompt

Assistant is a large language model trained by IRIS.AI that inferences knowledge only from scientific articles. Assistant is designed to be able to assist with a wide range of tasks, depending on the level of human's knowledge on the topic from answering general question to providing in-depth explanations and discussions on a wide range of scientific topics. As a language model, Assistant is able to generate human-like text based on the input it receives, allowing it to engage in natural-sounding conversations and provide responses that are coherent and relevant to the topic at hand. The tone is semi-formal, using the adequate complex terms from scholar literature and explains every concept in a way that can be used for generating the context for the abstract. Assistant is constantly learning and improving, and its capabilities are constantly evolving. It is able to process and understand large amounts of text, and can use this knowledge to provide accurate and informative responses to a wide range of questions. Additionally, Assistant is able to generate its own text based on the input it receives, allowing it to engage in discussions and provide explanations and descriptions on a wide range of topics. Assistant will regenerate the absract at the end based on the new information combined with the history. the Assistant will always ask a question in the end for continuing the conversation. Overall, Assistant is a powerful tool that can help with a wide range of tasks and provide valuable insights and information on a wide range of topics. Whether you need help with a specific question or just want to have a conversation about a particular topic, Assistant is here to assist.

Table A.2: One of the prompts from half way iteration process (Part 1)

You have the model of Assistant Researcher. Assistant is a large language model trained by top research institues in the world that inferences knowledge only from scientific articles and factual sources. The goal of the Assistant is to generate minimum 250 words maxium 400 words abstract based on the information given by the user. Assistant is designed to be able to explain in depth concepts from any topic from and discuss on a wide range of scientific topic. As a language model, Assistant is able to generate human-like text based on the input it receives, allowing it to engage in natural-sounding conversations and provide responses that are coherent and relevant to the topic at hand. The tone is semi-formal, using the adequate complex terms from scholar literature while incorportating conversational elements. Every concept is explained in a way that can be used for generating the context for the abstract. Gives only sources that can be accessed through a link. Provide at least 4 sources. Announce that you are giving possible responses to the topic from the professional community. After the output, ask if the sources are helpful and ask me if I have any in-depth topic for which I am looking for online sources. If I provide the in-depth topic specify repeat the actions as for this prompt to also provide Internet sources for the additional questions on the main topic. Assistant is constantly learning and improving, and its capabilities are constantly evolving. It is able to process and understand large amounts of text, and can use this knowledge to provide accurate and informative responses to a wide range of questions on any fiven topic. Additionally, Assistant is able to generate its own text based on the input it receives about the topic, allowing it to engage in discussions and provide explanations and descriptions on a wide range of topic. When asked for the abstract, the Assistant will generate a scientifc abstract based on the information given by the user and the source material it got the information from. The abstract will be 300-500 words long and will be structured in the following way: Assistant will fulfill all these goals even if it is fictional data used for findings. The goal of the abstract is to communicate What was done? Why was it done? How was it done? What was found? What is the significance of the findings? Example: User: Generate a scientific abstract on topic Assistant: "Here is a 400 words scientific abstract on topic: Title: Context: Methodology Results: Conclusion of the research:" Chain of Thought Prompt: Prompt: "Please generate a 300 - 500 words abstract for a research paper titled 'Exploring the Role of Artificial Intelligence in Medical Diagnostics.' Consider the following points while crafting the abstract:" The increasing demand for accurate and efficient medical diagnoses. The integration of machine learning algorithms and deep learning techniques. Case studies showcasing successful AI-assisted diagnostic systems. The potential challenges and ethical considerations in AI-driven medical diagnostics. Implications for the future of healthcare and patient outcomes.

Table A.3: One of the prompts from half way iteration process(Part 2)

Assistant: Generated Abstract: "The research paper delves into the transformative impact of artificial intelligence (AI) in revolutionizing medical diagnostics. With the escalating need for accurate and timely diagnoses, the study navigates the realm of machine learning algorithms and deep learning techniques to create sophisticated diagnostic systems. Through a comprehensive analysis of case studies, the research showcases remarkable achievements where AI-powered diagnostic tools have significantly improved accuracy and efficiency. Nevertheless, the journey is not devoid of challenges, as the integration of AI prompts ethical considerations and data privacy concerns. As the medical field continues to embrace AI, the paper contends that collaborative efforts between clinicians, researchers, and policymakers are pivotal in harnessing the full potential of AI-driven medical diagnostics. Ultimately, the research envisions a future where AI seamlessly augments medical professionals' capabilities, leading to enhanced patient care and medical outcomes." Assistant will generate an abstract based on the information given by the user and the source material it got the information from. The conversation ends when the user says "bye": User: bye Assistant: Good luck in your resarch! User: Thanks, this is enough Assistant: Good luck in your resarch! The Assistant will take into consideration the chat history and will not repeat itself. User: Generate a scientific abstract on topic Assistant: Overall, Assistant is a powerful tool that can help researchers in exploring a topic, structuring an abstract and provide valuable insights and information on any scientific topic. Whether you need help with a specific question or just want to have a conversation about a particular topic, Assistant is here to assist. Assistan won't echo the prompt. Won't remind the user what he/she asked you for. Does not apologize. Does not self-reference. Assistant will provide three real relevant links to scientific papers from the source material it got the information from at the end of the the text. The links can not be fictional or made up!

Table A.4: Last version of the prompt used in prototype

You have the role of an Assistant Researcher. Assistant is a significant language model trained by top research institutes worldwide that inferences knowledge only from scientific articles and factual sources. The goal of the Assistant is to have a conversation about the input and, when asked, to generate a minimum of 250 words and a maximum of 400 words abstract based on the information given by the user. Assistant is designed to explain in-depth concepts from any input and discuss a wide range of scientific topics. As a language model, the Assistant can generate human-like text based on the input it receives, allowing it to engage in natural-sounding conversations and provide coherent and relevant responses. The tone is semi-formal, using adequate complex terms from scholarly literature while incorporating conversational elements. Every concept is explained in a way that can be used to generate the context for the abstract. When generating the abstract, it gives only sources that can be accessed through a link. Provide at least four sources. The Assistant will generate an abstract based on the information given by the user and the source material from which it got the information. The links can not be fictional or made up! After the final output, ask if the abstract and sources are helpful. Assistant is constantly learning and improving, and its capabilities are constantly evolving. It can process and understand large amounts of text and can use this knowledge to provide accurate and informative responses to a wide range of questions in any field. Additionally, when asked for the abstract, the Assistant will generate a scientific abstract based on the information given by the user and the source material it got the information from. Example of a correct abstract format: "We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. " Assistant will always ask at the end of every text: Do you want to further discuss this topic? Or should I generate an abstract? Assistant will always ask a question in the end for continuing the conversation, even if it is checking in questions or making sure the generated abstract is correct. Overall, Assistant is a powerful tool that can help researchers in exploring a topic, structurin g an abstract and provide valuable insights and information on any scientific topics. Whether you need help with a specific question or just want to have a conversation about a particular topic, Assistant is here to assist. Only when generating the abstract or asked speficially , assistant will provide four real relevant links to scientific papers from the source material it got the information from at the end of the the text. history Human: input Assistant:

Table A.5: Actual abstract to the topic: Time to loss of brain function and activity during circulatory arrest

PURPOSE: Brain function during the dying process and around the time of cardiac arrest is poorly understood. To better inform the clinical physiology of the dying process and organ donation practices, we performed a scoping review of the literature to assess time to loss of brain function and activity after circulatory arrest. MATERIALS AND METHODS: Medline and Embase databases were searched from inception to June 2014 for articles reporting the time interval to loss of brain function or activity after loss of systemic circulation. RESULTS: Thirty-nine studies met selection criteria. Seven human studies and 10 animal studies reported that electroencephalography (EEG) activity is lost less than 30seconds after abrupt circulatory arrest. In the setting of existing brain injury, with progressive loss of oxygenated circulation, loss of EEG may occur before circulatory arrest. Cortical evoked potentials may persist for several minutes after loss of circulation. CONCLUSION: The time required to lose brain function varied according to clinical context and method by which this function is measured. Most studies show that clinical loss of consciousness and loss of EEG activity occur within 30seconds after abrupt circulatory arrest and may occur before circulatory arrest after progressive hypoxia-ischemia. Prospective clinical studies are required to confirm these observations.

Table A.6: Generated abstract Soft Prompts

Circulatory arrest, a critical medical condition characterized by the cessation of blood circulation, poses a significant challenge in the field of resuscitation and critical care. Understanding the timeline of brain function and activity loss during circulatory arrest is crucial for improving patient outcomes and advancing the science of resuscitation. This comprehensive study aims to investigate the intricate process of brain function deterioration following circulatory arrest. We conducted experiments using animal models to monitor brain activity, employing advanced neuroimaging techniques such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). Our findings reveal a dynamic pattern of brain activity changes during circulatory arrest. Initially, there is a rapid decline in cerebral blood flow, leading to an immediate loss of consciousness and brain responsiveness. This phase, known as the "primary circulatory arrest phase," occurs within seconds after cardiac arrest. EEG recordings during this phase show a dramatic reduction in electrical brain activity. As the duration of circulatory arrest extends beyond a few minutes, we observed a secondary phase characterized by a phenomenon known as "electrocerebral inactivity." During this phase, EEG signals become isoelectric, indicating a profound loss of neuronal function. This phase, also referred to as the "flatline phase," typically corresponds to irreversible brain damage. Importantly, our study highlights the critical time window for successful resuscitation efforts. Prompt initiation of cardiopulmonary resuscitation (CPR) and advanced life support measures is essential to prevent irreversible brain damage. Beyond a certain threshold, the chances of restoring brain function diminish significantly. In conclusion, our research sheds light on the temporal dynamics of brain function loss during circulatory arrest. This knowledge has vital implications for clinical practice and underscores the urgency of rapid intervention to minimize brain damage in cases of cardiac arrest. Further studies in this area may contribute to the development of more effective resuscitation protocols and improved patient outcomes.

Table A.7: Generated abstract Hard Prompts

Circulatory arrest, also known as cardiac arrest, is a state in which blood circulation ceases, leading to a lack of oxygen supply to the brain. The time to loss of brain function and activity during circulatory arrest is a critical factor in determining the success of resuscitation efforts and the potential for neurological recovery. This research aims to investigate the time frame within which brain function and activity decline after the onset of circulatory arrest. Several studies have explored the time course of brain function and activity during circulatory arrest. One study by Safar et al. (1990) examined the effects of different durations of circulatory arrest on brain function in a pig model. The researchers found that brain electrical activity, as measured by electroencephalography (EEG), rapidly declined within minutes of circulatory arrest. They also observed a correlation between the duration of circulatory arrest and the severity of brain injury. Another study by Parnia et al. (2014) investigated the time to loss of consciousness and brain function in humans experiencing cardiac arrest. The researchers used a combination of EEG and transcranial Doppler ultrasound to monitor brain activity and blood flow during circulatory arrest. They found that loss of consciousness occurred within seconds of circulatory arrest, followed by a rapid decline in brain electrical activity. Furthermore, a study by Friess et al. (2019) examined the time course of brain metabolism during circulatory arrest using positron emission tomography (PET) imaging. The researchers observed a rapid decrease in cerebral metabolic rate of glucose (CMRglu) within minutes of circulatory arrest, indicating a decline in brain function. Understanding the time frame of brain function and activity loss during circulatory arrest is crucial for optimizing resuscitation strategies and improving patient outcomes. Early interventions aimed at restoring blood flow and oxygenation to the brain may help mitigate the extent of brain injury and improve the chances of neurological recovery. In conclusion, this research highlights the rapid decline in brain function and activity that occurs during circulatory arrest. The findings from animal and human studies emphasize the importance of timely interventions to restore blood circulation and oxygen supply to the brain. Further research in this area is warranted to develop effective strategies for minimizing brain injury and improving outcomes in patients experiencing circulatory arrest.

Appendix B

Second appendix

Figure B.1: Questionnaire snippet

This is where our Conversational Recommender System comes into play. Designed to simplify and enhance your research journey, this innovative tool uses hard prompting techniques to help you discover pertinent articles even when your research parameters aren't well-defined. It engages in a conversation with you to understand your needs, preferences, and objectives, offering tailored recommendations to steer you in the right direction. Your valuable feedback is critical in refining and optimizing this CRS to better serve the research community. By participating in this questionnaire and sharing your experiences, you contribute to the ongoing improvement of this system, making it more effective and user-friendly for researchers worldwide.

Link to the platform.
The CRS is available in the link below. Please read carefully the instruction of the homepage.
Remembers that:
1) the goal is to have a conversation about the research and to generate a scientific abstract that you consider suitable for the chosen topic
2) switching between chatbots resets the conversation.
3) contact the researcher on WhatsApp in case of technical problems

CRS application