INFERENCE AND INTERPRETABILITY:

HOW NEURAL NLI MODELS PERFORM NATURAL LOGIC DEDUCTIONS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2023

By
Yulia Rozanova
Department of Computer Science

Contents

Abstract **6**

Declaration **8**

Copyright **9**

Acknowledgements **10**

1 Introduction **12**
1.1 Motivation 12
1.2 Background 15
1.2.1 Problem Definition 15
1.2.2 Problem Statement 20
1.3 Research Questions and Objectives 23
1.3.1 Building on Behavioural Observations 23
1.3.2 Observational Interpretability Study 24
1.3.3 Interventional Interpretability Study 25
1.4 Contributions 26
1.5 Thesis Outline 27
1.6 Publications 28

2 Structural Interpretability for NLP **29**
2.1 Introduction 29
2.2 Observational Methods: Representation Probing 31
2.2.1 Probing Tasks 32
2.2.2 Methodological Trends and Developments 35
2.3 Interventional Methods 39
2.4 Conclusion and Influence on Our Work 41

2.5 Scoping and Limitations 42

3 Supporting Context Monotonicity Abstraction 43
3.1 Introduction 43
3.2 Related Work 45
3.3 Experiments 47
3.3.1 Retraining NLI Models to Classify Context Monotonicity . . 48
3.3.2 Improving NLI Performance on Monotonicity Reasoning . . . 50
3.4 Discussion 53
3.5 Conclusion and Future Work 55
3.6 Scoping and Limitations 56

4 Decomposing Natural Logic Inferences in Neural NLI 57
4.1 Introduction 58
4.2 Related Work 59
4.3 NLI-*XY* Dataset 59
4.4 Experimental Setup 63
4.4.1 Model Choices 63
4.4.2 Probing Tasks 64
4.4.3 NLI Challenge Set Evaluations 66
4.4.4 Decomposed Error Analysis 67
4.5 Results and Discussion 67
4.5.1 Probing Results 67
4.5.2 Comparison to Challenge Set Performance 69
4.5.3 Qualitative Analyses 69
4.6 Conclusion 70
4.7 Scoping and Limitations 71

5 Interventional Probing in High Dimensions 77
5.1 Introduction 77
5.2 Interventional Probing 79
5.2.1 What Should it Tell Us? 80
5.2.2 The Amnesic Intervention 80
5.2.3 A Variation: The Mnestic Intervention 82
5.3 Experimental Setup 83
5.3.1 Dataset 83

5.3.2 NLI Models and Encoding . . . 85
5.3.3 Evaluation . . . 85
5.4 Results and Discussion . . . 86
5.4.1 Single Feature Amnesic Probing . . . 86
5.4.2 Multi Feature Amnesic Probing . . . 87
5.4.3 Mnestic Probing . . . 89
5.4.4 Control Comparison . . . 90
5.5 Qualitative Visualisations . . . 91
5.5.1 Visualisations of Unmodified Representations . . . 92
5.5.2 Visualisations of Interventions . . . 92
5.6 Related Work . . . 93
5.7 Conclusion and Future Work . . . 95
5.8 Scoping and Limitations . . . 96
5.9 Expanded Amnesic Intervention Results . . . 97

6 Causal Effects of Natural Logic Features 100
6.1 Introduction . . . 101
6.2 Problem Formulation . . . 102
6.2.1 A Structured NLI Subtask . . . 102
6.2.2 The Causal Structure of Model Decision-Making . . . 103
6.3 Estimating the Causal Effects . . . 105
6.3.1 Interventions for Calculating TCE and DCE . . . 106
6.4 Experimental Setup . . . 109
6.4.1 Data and Interventions . . . 109
6.4.2 Model Choice and Benchmark Comparison . . . 110
6.5 Results and Discussion . . . 111
6.5.1 Causal Effect of Inserted Word Pairs . . . 111
6.5.2 Causal Effect of Contexts . . . 113
6.5.3 Benchmark Scores and Causal Effects . . . 114
6.6 Related Work . . . 115
6.7 Conclusion . . . 116
6.8 Scoping and Limitations . . . 117

7 Conclusion 118
7.1 Summary and Conclusion . . . 118
7.2 Opportunities for Future Work . . . 120

7.3 Longevity of This Work 122
7.4 Ethical Implications 122

A Code, Data and Models **142**
A.1 Supporting Context Monotonicity Abstraction in Neural NLI 142

Abstract

Example-based learning for Natural Language Understanding (NLU) tasks has been a long-standing goal of Artificial Intelligence (AI) and has seen major success as machine learning methods, architecture capacities and the scale of data processing capabilities have improved in recent decades.

However, training large, opaque models on very high-level objectives such as Natural Language Inference (NLI) raises fundamental questions about whether appropriate reasoning strategies have been learnt by a given model. In fact, much work has highlighted the emergence of spurious heuristics which aid NLI model performance in unexpected ways, rather than following theoretically expected systematic reasoning routes using appropriate properties.

Research on *model interpretability* has been providing rapidly maturing methodologies which may shed light on the linguistic features and abstract properties captured within the representations of trained models, as well as their comparative effects on model predictions.

This thesis isolates a structured subtask of NLI based on *natural logic* as a framework for applying and developing interpretability methods with the end goal of better assessing the reasoning capabilities of NLI models. In particular, we model entailment examples which are single-step natural logic deductions relying on exactly two abstract semantic features: hierarchical *concept relations* and the *monotonicity* of a natural language context. Responding to behavioural observations of NLI model limitations, we turn to both observational and interventional interpretability methods to analyze competitive NLI models' abilities to perform natural logic deductions and diagnose failure patterns at a finer granularity.

Overall, the scientific contributions present in this thesis can be summarised as follows:

1. A study of existing strategies for improvement of natural logic handling in neural NLI, together with evaluations on previously established monotonicity

reasoning evaluation sets. This is complemented with an introduction of a context monotonicity prediction task for a transfer learning improvement strategy for NLI models. The study reveals that there is not much additional performance to gain from this strategy over existing improvement strategies. Overall, the experimental results suggest that enough information may already be gleaned from fine-tuning on the HELP dataset to support internal latent modelling of the context monotonicity feature.

2. The construction of the NLI-*XY* Dataset, which is suitable for interpretability methods and qualitative error analyses.

3. An extensive *probing* study to determine the representation of the two semantic features relevant to natural logic, context monotonicity and concept inclusion relations. Specifically, we draw comparisons between NLI models before and after improvement strategies for natural logic handling, showing SOTA models fail to capture the monotonicity feature, while slightly fine-tuned models demonstrate strong emergence of this feature after some more balanced training.

4. Qualitative analyses in the form of *visualised projections* and informative *error breakdowns* which further bolster the argument that it is poor *context monotonicity* modelling that is a bottleneck for strong out-of-the-box NLI models' capacity for correctly identifying valid natural logic deductions.

5. An application of the *amnesic probing* methodology of Elazar, Ravfogel, Jacovi, *et al.* [1] which follow on from our previous structural findings, as well as a discussion of the limitations of this methodology in this context.

6. The introduction of an alternative interventional probing method which we call *mnestic* probing, a variation of *amnesic* probing, which yields more informative results in our problem setting.

7. A study of direct and indirect causal effects related to context monotonicity and concept inclusion relations, following the framework of Stolfo, Jin, Shridhar, *et al.* [2] in using these measurements as indicators of *robustness* and *sensitivity*.

Declaration

i. No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

ii. The material presented in this thesis represents the candidate's own work except where stated otherwise.

Copyright

Acknowledgements

I am unboundedly grateful to my supervisor, Andre Freitas, for the incredible opportunity this PhD journey has been. He has worked tirelessly to build a strong and highly collaborative research group, and his ability to continue to do so despite the setbacks of the COVID-19 pandemic speaks wonders for his community building skills. This work would not have been possible without the scaffolding he has set in place, and his dedicated support at every stage of the PhD. I am also grateful to my co-supervisor Uli Sattler: she is an inspirational force of nature, and I only wish I had taken more advantage of her valuable conversations and sharp intellect.

I would like to thank my examiners Marco Kuhlmann and Ian Pratt-Hartmann for their thorough review of my thesis and the stimulating conversations during the viva. To have them engage with my work with the level of interest, insight and rigour that they have shown has been both a privilege and a delight.

I have been lucky to have my time in Andre's research group overlap with three particularly wonderful fellow students: Deborah Ferreira, Mokanarangan Thayaparan and Marco Valentino. It has been a joy to find many opportunities of collaboration with them, and even a greater joy to find such precious friends. I cannot thank them enough for their companionship, discussions, and most importantly, occasional nudges of encouragement which have pushed me through the paralysis of self-doubt. A special thanks to Marco for his supervisory input towards the end of my PhD.

I am grateful to be supported by the University of Manchester through their award of the International Student Fee Waiver. The greater community at the University of Manchester has been a rich source of collaboration opportunities, social events and fantastic individuals to befriend. A big thanks to the Enncore research group, especially Lucas Cordeiro, Edoardo Manino and Danilo Carvalho, for the opportunity to be involved in some cross-disciplinary projects. I would like to thank everyone I have encountered at the University of Manchester who has enriched my PhD experience in the department: in particular, Jake, Ruba, Martina, Maddie, Connor, Oskar, Hanadi,

Mingyang, Mario, and lastly, Mauricio, with whom it was always a joy to dance, even in the department cafe. A few more individuals in Manchester have been a great part of my support network and my life in this city: our comrade Alber Santos, and finally Catherine Sharrock, Sarah Featherstone, Monty, Hector and the extended Sharrock/Featherstone family.

The journey to the PhD was already strongly supported by my community at the University of Stellenbosch, South Africa, where I was lucky to do my masters and undergraduate studies. For this, I would like to thank both the mathematics department (especially Ingrid Rewitzky and Zurab Janelidze), and various members of the engineering department whose help has been crucial in my transition from mathematics to computer science (in particular, Ben Herbst, Steve Kroon and Willie Brink). I also cannot state strongly enough the positive impact that the *Deep Learning Indaba* has had on this transition. Especially during the pandemic, it has been grand to stay in touch with my Stellenbosch comrades in mathematics and video games: thank you to Sarah Selkirk, Brandon Laing and Luke Vorhies.

A special thanks to my partner Patrick Clarke, whose support has been invaluable, especially through the most challenging moments in the final stages of my PhD. You have always helped me to stay joyful, relax and keep track of the important objectives.

Finally, I am extremely grateful for the support and impact of my parents, Dr Andre Rozanov and Tatiana Tarassova. Especially in the world of academia, the implicit advantages afforded to children of parents who have gone through tertiary education themselves cannot be overstated. I have been incredibly lucky to grow up absorbing their wisdom, critical thinking skills, mastery of language, and most importantly, an attitude of playful curiousity. I cannot thank them enough for their love, encouragement and always delighting in even my most mundane successes.

Chapter 1

Introduction

1.1 Motivation

Example-based learning for Natural Language Understanding (NLU) tasks such as Question Answering (QA) and Natural Language Inference (NLI) has been a long-standing goal of Artificial Intelligence (AI) and has seen major success as machine learning methods and the scale of data processing capabilities have improved in recent decades. In particular, so much success has come from the advent of transformer-based models [3]–[5] that these have come to dominate the NLP landscape, topping the leaderboards of NLU benchmark datasets such as GLUE [6].

However, training large, opaque models on high-level objectives such as NLI raises questions about whether we can trust that appropriate reasoning strategies have been learnt by a given model. In fact, much work has highlighted the emergence of spurious heuristics which aid NLI model performance in unexpected ways [7]–[9], rather than following theoretically expected systematic reasoning routes using relevant properties.

Research on *model interpretability* has provided rapidly maturing methodologies which may shed light on the linguistic features and abstract properties captured within the representations of trained models, as well as their comparative effects on model predictions. For example, works such as Tenney, Das, and Pavlick [10] and Hewitt and Manning [11] use probing techniques to demonstrate that the representations of transformer-based language models capture linguistic information which is associated with the features used in traditional NLP pipelines (including *syntactic* information such as part-of-speech tags and *semantic* information such as semantic role labels). To increase our understanding and identify areas of potential improvement in the reasoning strategies of pretrained NLI models, we would like to be able to identify

when models use expected features in a principled and systematic way, and when they are falling short.

This thesis isolates a structured subtask of NLI based on *natural logic* as a framework for applying and developing interpretability methods with the end goal of better assessing the reasoning capabilities of NLI models. In particular, we model entailment examples which are single-step substitutions where the entailment label depends on exactly two abstract semantic features: hierarchical *concept relations* and the *monotonicity* property of a natural language context, as exemplified in figure 1.1 and fully desribed in section 1.2.1.

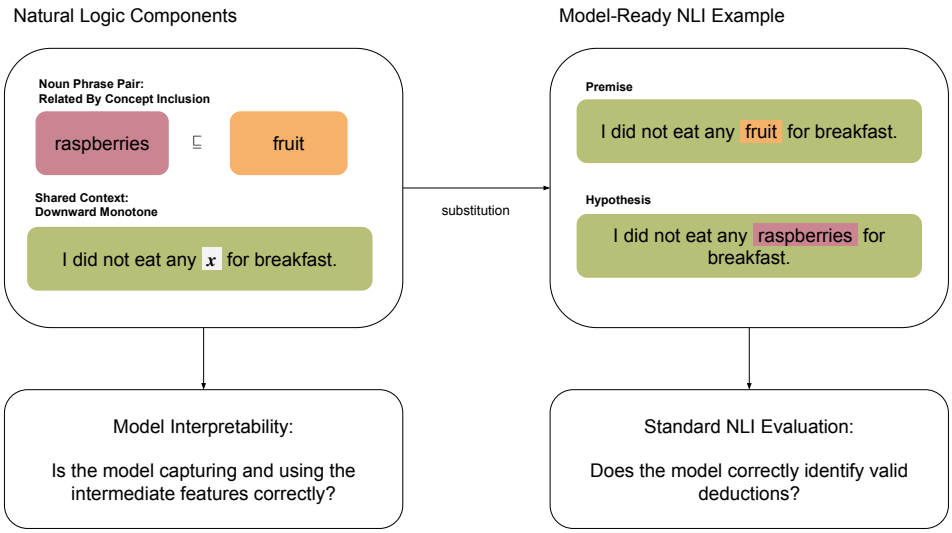


Figure 1.1: An example of the type of NLI problem that will be considered in this work. Whether such examples are to be labelled as *entailment* or *non-entailment* depends on the relation between the substituted phrases and the *monotonicity* property of the shared context: the presence or absence these features in NLI models' encodings will be examined in the interpretability experiments in later chapters.

Responding to behavioural observations of NLI model limitations, we turn to both observational and interventional interpretability methods to analyze competitive NLI models' abilities to perform natural logic deductions and diagnose failure patterns at a finer granularity.

Natural Logic This thesis focuses on the setting of *natural logic* as a framework for advancing research in systematic inferential capabilities of NLI models. The motivation for isolating natural logic phenomena within the greater areas of NLI and model interpretability is two-fold:

- There is a solid groundwork of behavioural evaluation in the natural logic setting [12]–[15] which has identified insufficiencies even in well-performing NLI models (with respect to popular benchmarks). However, application of structural interpretability methods [16] has so far been lacking. In part, this is due to the somewhat more complex theoretical formulations of natural logic, which we present in a more abstracted way which is better suited to interpretability studies. On the other hand, it is the *recent improvement stratgies* for NLI models' natural logic handling capabilities that have made it more likely that we may observe structural emergence of natural logic features within models.

- Studies in model interpretability are overwhelmingly centered on task-agnostic linguistic features whose influence on the model objective (e.g. masked language modelling) is too obscure and diluted in order to draw up structured expectations. In the absence of expectations, it is both difficult to validate new interpretability methods, to extend interpretability results to arguments about *reasoning capabilites*, and to transition from *observational* to *interventional* methods. We see natural logic as a fruitful setting in this regard, as it has just enough structure to create relatively simple causal expectations.

Transformer-Based NLI Models Shortly before the commencement of this project, the release of the BERT model [17] signalled the advent of a new era of transformer-based solutions to NLP tasks. Transformers differed from previous language model architectures in their integration of an attention mechanism [3] which allowed for rich *contextual* representations for all sequence tokens, as well as in their utilization of a new *masked language modelling* training objective. After pre-training, transformer models can be *fine-tuned* on a secondary task, such as NLI .

The introduction of these architectures was followed by an expansion in NLI models which greatly surpassed the performance of previous models on well-known NLI task benchmarks such as MNLI [18] and SNLI [19]. A particularly successful model was built on RoBERTa [5] (a BERT variation which uses a uses a bytepair encoding scheme and is trained for more epochs), which is one of the core models we include throughout this work (namely, roberta-large-mnli). Furthermore, a trend in *model interpretability* has increasingly shown the capabilities of transformers to model complex linguistic features [11] which they are not explicitly trained to classify. Despite this success, transformer-based NLI models were subjected to rigorous scrutiny which highlighted the exploitation of spurious patterns that did not mimic the desired principled behaviours

of inference models [7].

Our motivation for focusing on transformer-based NLI models follows on from various observations of their limitations, especially with respect to natural logic phenomena [12], [14], as well as interest in the burgeoning area of model interpretability, which has been introducing new methodologies for slicing open and *probing* the representations of these pragmatically performant but demonstrably flawed and opaque models.

1.2 Background

1.2.1 Problem Definition

Natural Language Inference (NLI) is a classification task aimed at identifying entailment relations between sentence pairs. Given a *premise p* and a *hypothesis h* (both natural language sentences), the simplest form of the NLI task is to provide a true or false classification as to whether *p* entails *h* (having read *p*, a human reader could conclude that the truth of *h* can be inferred from the truth of *p*). A common three class variation of the problem is to determine whether *p* entails *h*, *p* and *h contradict* each other, or neither relationship holds (the *neutral* label).

The structure of positive entailment examples can vary broadly:

(1) *"The lawyer knew that the judges shouted."* implies *"The judges shouted."*. [7]

(2) *"Mary used her workstation."* implies *"Mary has a workstation."*. [20]

(3) *"An Irishman won the Nobel prize for literature."* implies *"An Irishman won a Nobel prize."* [14]

A positive entailment label could, for example, be a result of claims in the hypothesis *h* that are explicitly or implicitly contained as a subclaim in the premise *p* (such as in example (1)), or it could be dependent on intra-sentential anaphora resolution (such as in example (2)), or it may depend on the inclusion relation between concepts (as in example (3), where the concept "the Nobel prize for literature" is semantically contained in the concept "a Nobel prize").

In the earlier linguistic tradition of building (or theorising about) *symbolic* natural language reasoning systems, authors have described a systematic scoped subset of natural language inference problems which is refered to as *natural logic* [21]–[23]: this can be presented as a subtask of the modern the NLI task. Broadly speaking, it

relates to how we may reason about concepts related by a semantic *concept inclusion* relation (like in example (3)), while taking into account the logical effect of functional words (such as negations and generalised quantifiers). For our purposes, we choose a presentation of natural logic that may be seen as a form of *substitutional reasoning*.

Consider the following example of a single step natural logic inference, in which a noun phrase is substituted for another, yielding a sentence which is entailed by the first:

(4) Premise: I did not eat any **fruit** for breakfast.
Hypothesis: I did not eat any **raspberries** for breakfast.
NLI Label: Entailment

Here, the word *fruit* is substituted with a more specific concept, *raspberries*. The hyponym/hypernym pair (raspberries, fruit) exemplifies a more general relation which we will refer to as the *concept inclusion* relation $\sqsubseteq$, (and dually, *reverse concept inclusion* $\sqsupseteq$), which are analagous to set-theoretic inclusion relations ($\subseteq, \supseteq$). For clarity, we may also use the symbol # to emphasize an unrelated pair. As we exemplify in table 1.1, we consider the relation to apply to arbitrarily complex noun phrases such as noun phrases modified with adjectives and prepositional phrases), which distinguishes it from the lexical relations of *hypernymy/hyponymy*.

	x	y
$\sqsubseteq$	raspberries	fruit
	brown sugar	sugar
	dogs with hats	dogs
#	computer	potato

Table 1.1: Examples which demonstrate the *concept inclusion relation* $\sqsubseteq$.

We refer to the shared part of the sentence "I did not eat any ___ for breakfast" as the shared *context*. We will refer to it by a functional symbol f, treating it as a function that takes a noun phrase argument. In this work, we consider only the subset of NLI examples where the premise–hypothesis pairs have the structure $(p = f(x), h = f(y))$, where f is a shared context and (x, y) is a pair of distinct noun phrases with a known concept inclusion relation ($\sqsubseteq, \sqsupseteq$ or #), where x and y are each *inserted* into the context f.

For examples structured in this way, the final entailment label depends on two properties: the relation between the substituted concepts, and the nature of the linguistic context surrounding the substituted phrase: in particular, a property of the context called its *monotonicity*.

			Example 1	Example 2
Components	**Context**	f	I did not eat any ___ for breakfast	I ate some ___ for breakfast
	Context Monotonicity		down $\downarrow$	up $\uparrow$
	Concept Pair	(x, y)	(fruit, raspberries)	(raspberries, fruit)
	Concept Relation		$\sqsupseteq$	$\sqsubseteq$
NLI Example	**Premise**	$f(x)$	I did not eat any fruit for breakfast	I ate some raspberries for breakfast
	Hypothesis	$f(y)$	I did not eat any raspberries for breakfast	I ate some fruit for breakfast
	NLI Label		Entailment	Entailment

Table 1.2: Two examples following the structure of NLI problems we consider in this work. Example 1 features a downward monotone context, while example 2 features an upward monotone one.

Context Monotonicity We say that the context f is *upward monotone* ($\uparrow$) if, given a sentence $f(x)$, we may substitute the noun phrase x for a noun phrase y which is more *general* (in particular, where $x \sqsubseteq y$) and the resulting sentence $f(y)$ is entailed by the sentence $f(x)$. On the other hand, we say that f is *downward monotone* ($\downarrow$) if, given a sentence $f(x)$, we may substitute the noun phrase x for a noun phrase y which is more *specific* (in particular, where $x \sqsupseteq y$) and the resulting sentence $f(y)$ is entailed by the sentence $f(y)$. In table 1.2, we give two NLI examples with a positive entailment label: one of which features an upward monotone context, one which features a downward monotone one. This table also illustrates how our described components (a shared context and a pair of noun phrases) combine into a premise and hypothesis pair, which is the standard NLI model input.

Linguistically, the monotonicity of the context is usually influenced by words such as negations (for example "not", "no") or generalised quantifiers (such as "every"). Much of the linguistic work in natural logic has been the characterization of the monotonicity of various linguistic operators. We include some examples of downward monotone operators in table 1.3 from Yanaka, Mineshima, Bekki, *et al.* [13]: this demonstrates the diversity of downward monotone operators, showing that they are not limited to certain parts of speech.

The monotonicity of a noun phrase "slot" in a context is determined by the monotonicity of all of the operators in whose scope it falls: an odd number of downward monotone operators yields a downward monotone context, while an even number of downward monotone operators results in an upward monotone context. For example, the context "every ___ barks" is downward monotone, while "not every ___ barks" is upward

Category	Examples
determiners	*every, all, any, few, no*
negation	*not, n't, never*
verbs	*deny, prohibit, avoid*
nouns	*absence of, lack of, prohibition*
adverbs	*scarcely, hardly, rarely, seldom*
prepositions	*without, except, but*
conditionals	*if, when, in case that, provided that, unless*

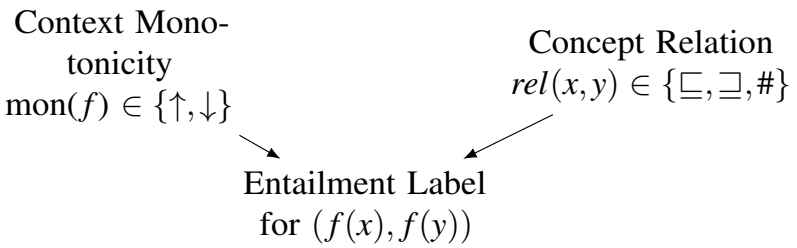
Table 1.3: Examples of downward monotone operators.

$Mon(f)$ \ $Rel(x,y)$	$\sqsubseteq$	$\sqsupseteq$	#
$\uparrow$	Entailment	Non-Entailment	Non-Entailment
$\downarrow$	Non-Entailment	Entailment	Non-Entailment

Table 1.4: The entailment gold labels as a function of two semantic features: the context montonicity (Mon(f)) and the relation (Rel(x,y)) of the inserted word pair.

monotone. For an in-depth description of the monotonicty profiles of specific terms and operators, see works such as Sanchez [21] or MacCartney and Manning [23].

For the set of NLI examples of the $p = f(x)$, $h = f(y)$ structure, the *context monotonicity* and the *concept inclusion relation* jointly determine the final gold entailment label, as illustrated in the diagram below and following the schema in table 1.4.



Our central enquiry is whether existing trained NLI models can correctly detect the entailment labels of such NLI examples, and whether they are doing so by determining, storing and using information about the context monotonicity and the insertion pair relation in order to reach a decision about the final entailment label. If it is not able to do so, we would not expect an NLI model to be able to generalise well within this class of problems. We will later turn to methods in *model interpretability* to answer this quesion.

For an NLI model, there is a non-trivial amount of linguistic knowledge required to determine either of these intermediate features. Correctly classifying the relation between concepts requires not only a latent lexical hierarchy (hypernymy/hyponymy)

but the capability to interpret the relations between composed noun phrases which include modifiers and prepositional phrases (such as "dogs" and "dogs with hats"). Context monotonicity is a complex feature, which can further be decomposed into the monotonicity profiles of the linguistic operators which give rise to it. In this work, we are only interested in whether NLI models reflect the coarsest decision mechanism of how the final monotonicity label and the concept relation give rise to the entailment label. This is because it is simple enough to hypothesise and test how this reasoning strategy may be reflected in NLI models' internal representations, and to define a simple causal diagram which may be used to guide interventional experiments. In particular, our coarse segmentation of the monotonicity feature makes the interpretability work in chapters 5 and 6 possible.

As such, we do not focus on the decomposition of the monotonicity label in our problem definition or our experiments, but point to works such as Hu and Moss [24], Hu, Chen, Richardson, *et al.* [25] and MacCartney and Manning [23], which provide different formalisms of the compositional monotonicity phenomenon to the end of creating symbolic natural logic inference engines. We mention especially the *ccg2mono* system from Hu and Moss [24] which assigns a "polarity" marking to every word in a sentence, taking into account the monotonicity of relevant operators. This polarity marking indicates when a word may be replaced with one with a more general or more specific meaning, yielding an entailed sentence. This is equivalent to our notion of the monotonicity of that word's context, and it is owing to this system that we have the context monotonicity labels used in our experimental work via the HELP [14] and MED [13] datasets which have used it as part of their data creation process.

A Note on Notation When describing concept pairs (x, y) in ensuing chapters (especially when they represent existing publications), we may at times refer to them as *insertion pairs* or *word/term pairs* either to reflect the practicalities of the dataset construction, or for simplicity (but we note that the concepts are not necessarily single words). The concept inclusion relation $\sqsubseteq$ is thus sometimes refered to as the *insertion pair* relation, the *word pair* relation or the *lexical* relation.

Across the literature, the term *natural logic* is also described as *monotonicity logic* or *monotonicity reasoning*. In Richardson, Hu, Moss, *et al.* [12], it is described as a *fragment* of the NLI task, while in this work we more commonly refer to it as a *subtask*. As we are only dealing with a single noun-phrase substitution at a time, our formalized examples may themselves be seen as a subtask of a more general idea of

natural logic inferences, as in MacCartney and Manning [23] and Hu and Moss [24]. However, since we are often dealing with compound noun phrases, these can technically still be seen as multi-step natural logic deductions, so they are not dramatically less complex. Throughout the ensuing chapters, we use the terms *natural logic*, *monotonicity reasoning* and *natural logic subtask* interchangeably to describe our fragment of NLI examples, but note that this may not coincide exactly with these terms descriptions in related works.

1.2.2 Problem Statement

In this section, we describe a set of challenges and limitations at the intersection of neural NLI and model interpretability which directly motivate the research questions and subsequent work presented in this thesis.

1.2.2.1 Behavioural Limitations of NLI Models

The advent of transformer-based [3] language models such as BERT [4] and RoBERTa [5] has resulted in impressively high-performing NLI systems with respect to popular benchmark datasets such as SNLI [19] and MNLI [18].

However, these models are opaque in their reasoning and are primed to exploit spurious correlations and artifacts in the training data [7], [9]. For both transformer-based NLI models and earlier neural architectures, there is a rich research landscape of investigations into their behaviour and treatment of targeted phenomena [7]–[9], [12], [13], [26]–[30].

Some works have identified particular heuristics adopted by NLI models, such as a hypothesis-only bias [9], [28], over-reliance on lexical overlaps [7], [26] and subsequences [31], or the presence of specific words: for example, negation words are strongly correlated with a "contradiction" label [9], [31].

Especially in light of the observed issues with negation handling, variations of the natural logic task have received targeted evaluation treatment in several works, namely Richardson, Hu, Moss, *et al.* [12], Yanaka, Mineshima, Bekki, *et al.* [13], [14], and Geiger, Richardson, and Potts [15]. In particular, Yanaka, Mineshima, Bekki, *et al.* [13] and [15] have introduced monotonicity reasoning evaluation datasets (MED and MoNLI) which revealed a tendency of top-performing out-of-the-box-NLI models to consistently underperform on a subset of examples which require *downward monotone*

reasoning. This hints towards a lack of implicit monotonicity modelling in state-of-the-art NLI models, which would complement previous observations about negation handling.

There are, however, existing improvement strategies: both Richardson, Hu, Moss, *et al.* [12] and [14] demonstrate that with additional fine-tuning on the NLI objective with a balanced set of montonicity reasoning examples, performance on the targeted evaluation sets can be made to improve.

In summary, we note the following observations:

1. NLI models are seen to handle negations poorly.
2. NLI models often fail to classify entailment relations when faced with a downward monotone context, suggesting that the concept of *monotonicity* may be absent from the information captured and represented by models.
3. Their performance on natural logic challenge tasks can be improved by fine-tuning on an additional set of NLI examples with balanced monotonicity representation.

1.2.2.2 Lack of Structural Interpretability

Behavioural studies give rise to hypotheses about the features which are being taken into account by models, but *structural interpretability* studies are needed to make direct observations of model internals to determine if and where certain information is captured.

As such, the observations made about natural logic handling in NLI models allows for the hypothesis that state-of-the-art models fail to model monotonicity while "innoculated" versions of the same models might be doing so, but work still needs to be done in order to *quantify* how well the respective intermediate features are captured structurally in the models. Some steps have indeed been made in this direction: for example, Geiger, Richardson, and Potts [15] provide their own natural logic formalism (limited to the "not" downward monotone operator) which allows for some interpretability, and they perform a probing study which demonstrates the presence of *lexical hierarchical relations* in the representations of NLI models. They do not, however, run experiments on detectability of context monotonicity

This problem space can also be seen as a useful addition to the field of NLP interpretability in general: we outline in section 2.2.1 that there has not been much work on interpretability for *task-specific* features of models fine-tuned for specific natural language understanding tasks (as opposed to foundational pretrained language models).

In summary, we highlight these general limitations:

1. The majority of interpretability work in NLP focuses on task-agnostic linguistic features, limiting the capacity for hypothesising about model reasoning patterns.

2. Behavioural studies are suggesting that state-of-the-art models fail to capture proper monotonicity reasoning strategies, but there is no structural evidence that this is due to a lack of the context monotonicity feature in model representations.

3. Previous structural interpretability work shows that NLI models show a strong ability to capture *lexical relations* (a specific case of our *concept inclusion* relations) in their intermediate representations, but context monotonicity has not yet been structurally observed in the context of NLI models.

1.2.2.3 Interventions and Causal Influence

Our observations in chapter 4 demonstrate that certain models are indeed capturing context monotonicity and concept relations, but strong arguments have been made that high probing performance only allows us to deduce a correlational connection between model behaviour and the model internals [32]. In fact, Ravichander, Belinkov, and Hovy [33] find that even task-irrelevant features can be learnt to be predicted with a high accuracy by external probes. As such, we cannot yet make the leap that the NLI models which are seen to encode monotonicity in their representations are actually *using* this feature in the way they should. Researchers advocate for the use of *interventional* studies to support the stronger claim of feature use.

Elazar, Ravfogel, Jacovi, *et al.* [1] present an interventional method that is tightly linked with probing methodology: *amnesic* probing. This strategy relies on *projecting away* the representation subspace most linked with high probing performance for a given task, resulting in a representation from which the target feature has effectively been "removed". The modified representations are fed once again into the final task classifier, and a drop in performance is taken to suggest that the target feature was a necessary component for correct classification. This approach is our first port of call in chapter 5, but we discover some curious limitations: amnesic probing of the *gold label* of the final task fails to result in a performance drop, and similarly for our features of interest. As well-suited as the amnesic probing paradigm is for our setting, its inefficacy leaves a gap for a more informative probing-based interpretability method. We also note that the previous tasks to which amnesic probing has been applied do not come

with clear-cut *expectations* for how the features should impact model behaviour, while the natural logic setting does - this is an advantage that our work brings to the greater interventional interpretability landscape.

Lastly, an argument that a given feature is being used by a model should draw on causal modelling and present causal effect measures, which our applied interventional probing methods do not do. However, there is the potential to apply input-level interventional methods (such as Stolfo, Jin, Shridhar, *et al.* [2]) which are used to estimate causal influence. While causal effect estimations have been applied to problems such as examining the influence of gender features in text [34] and syntax [35], there does not yet appear to be such work applied to NLI.

In summary, we highlight the following gaps:

1. Our early interpretability studies follow methods that can only bring about *correlational* observations: for stronger evidence of causal impact, we need to turn to *interventional* studies.

2. Despite strong probing results, our application of *amnesic probing* results in some contradictory results, indicating that the method is insufficient for interventions in our setting.

3. The strongest form of causal argument relies on the measurement of *causal effects*, which has not yet been employed in an NLI setting.

1.3 Research Questions and Objectives

Our overarching concern is the reasoning behaviour of NLI models. The natural logic setting offers sufficient structure and simplicity to apply and develop interpretability methods that can shed light on whether models are capturing and using well-understood task-specific features. This thesis aims to answer the question:

RQ 0: *How well do existing NLI models perform natural logic deductions, and to what extent are they implicitly modelling context monotonicity and concept inclusion relations to do so in a systematic way?*

1.3.1 Building on Behavioural Observations

The first publication presented in this thesis (chapter 3) follows up on the observations in section 1.2.2.1. As a first step, we consider the hypothesis that poor context monotonicity

modelling is a bottleneck for natural logic performance, and ask:

> **RQ 1:** *How much does fine-tuning on the HELP dataset improve NLI models' performance on existing natural logic evaluation datasets? Would a secondary transfer-learning task based on the prediction of context monotonicity result in an improvement in overall evaluation scores?*

The results of our experiments showed that such a secondary training objective did not significantly improve upon the natural logic performance gains observed in [14]. We hypothesise that existing fine-tuning strategies may have already resulted in context monotonicity as an emergent latent feature within the model. To investigate this idea, we turn to the field of *model interpretability* to dig deeper and make comparisons between existing models and training regimes.

In chapter 2 we outline the relevant background in model interpretability which we have considered, with the end of answering the question:

> **RQ 2:** *Which interpretability methods are best-suited for our interest in detecting* emergent *intermediate features?*

With our interpretability goals in mind, we ask:

> **RQ 3:** *How can we construct a natural logic dataset that is suitable for both targeted evaluation and interpretability?*

We address **RQ 3** by introducing a compositional dataset (NLI-*XY*) in chapter 4.

1.3.2 Observational Interpretability Study

Formalising the NLI-*XY* dataset has allowed for structural interpretability work that aims to disentangle the representation of context monotonicity and concept inclusion relations. As an initial step, we observed that visualisations (available in chapter 4) of the improved NLI models' projected vector representations show a strong clustering behaviour which distinguishes between downward and upward monotone contexts, which is notably absent in the baseline state-of-the-art models. This supported the suggestion that a notion of concept monotonicity was emergent after fine-tuning on the HELP dataset introduced in Yanaka, Mineshima, Bekki, *et al.* [14], but the goal of the work in chapter 4 is to further support this with a systematic and quantitative structural interpretability study. Hence, in that work, we ask:

> **RQ 4:** *Are the intermediate features of context monotonicity and concept inclusion relations emergent in the internal representations of NLI models*

which perform better at natural logic tasks? Can we provide comparative quantitative evidence?

We address this by carrying out an extensive *probing* study, with the choice of methodology supported in chapter 2. Furthermore, the compositonal structure of our dataset allows for informative qualitative error breakdowns (presented as heatmaps in chapter 4), which, in combination with the probing and visualisation work, help us to answer the question:

RQ 5: *Which features are responsible for errors in poorer-performing models?*

1.3.3 Interventional Interpretability Study

In section 1.2.2.3, we discussed how our probing work is limited in the conclusions we are able to draw. We build on the existing observations by asking:

RQ 6: *What can structural interventional methods tell us about the usefulness of the identified representations for the NLI task?*

Our first attempt at addressing **RQ 6** focuses on the existing *amnesic probing* methodology, which we apply and present in chapter 5. As we will discuss in more detail, the application of amnesic probing yields unexpected (and even contradictory) results. This led us to consider:

RQ 7: *How can we devise an alternative interventional interpretability method that is still informative in the high-dimensional situations where amnesic probing fails?*

We introduce the alternative methodology of *mnestic probing* in chapter 5 to address **RQ 7**.

Despite the interventional nature of these techniques, it is still our aim to incorporate *causal* measures into our study, as the strongest level of interventional argument for the influence of our features of influence. In our final chapter, we ask:

RQ 8: *What can causal effect measures from interventional experiments tell us about NLI models' robustness and sensitivity to different types of intermediate feature changes?*

We address **RQ 8** by following the structure of Stolfo, Jin, Shridhar, *et al.* [2] to design a re-arrangement of our dataset into *intervention* sets which allow us to calculate certain causal effect measures. As well as providing a proxy measure of robustness and sensitivity, this adds a final causal layer to our argument about the emergence and

use (or lack thereof) of context monotonicty and concept inclusion relation features in neural NLI models.

1.4 Contributions

The contributions introduced in the works comprising this thesis may be summarised as follows:

C 1. A study of existing strategies for improvement of natural logic handling in neural NLI, together with evaluations on previously established monotonicity reasoning evaluation sets. This is complemented with an introduction of a context monotonicity prediction task for a transfer learning improvement strategy for NLI models. The study reveals that there is not much additional performance to gain from this strategy over existing improvement strategies. Overall, the experimental results suggest that enough information may already be gleaned from fine-tuning on the HELP dataset to support internal latent modelling of the context monotonicity feature.

C 2. The construction of the NLI-*XY* Dataset, which is suitable for interpretability methods and qualitative error analyses.

C 3. An extensive *probing* study to determine the representation of the two semantic features relevant to natural logic, context monotonicity and concept inclusion relations. Specifically, we draw comparisons between NLI models before and after improvement strategies for natural logic handling, showing SOTA models fail to capture the monotonicity feature, while slightly fine-tuned models demonstrate strong emergence of this feature after some more balanced training.

C 4. Qualitative analyses in the form of *visualised projections* and informative *error breakdowns* which further bolster the argument that it is poor *context monotonicity* modelling that is a bottleneck for strong out-of-the-box NLI models' capacity for correctly identifying valid natural logic deductions.

C 5. An application of the *amnesic probing* methodology of Elazar, Ravfogel, Jacovi, *et al.* [1] which follow on from our previous structural findings, as well as a discussion of the limitations of this methodology in this context.

C 6. The introduction of an alternative interventional probing method which we call *mnestic* probing, a variation of *amnesic* probing, which yields more informative results in our problem setting.

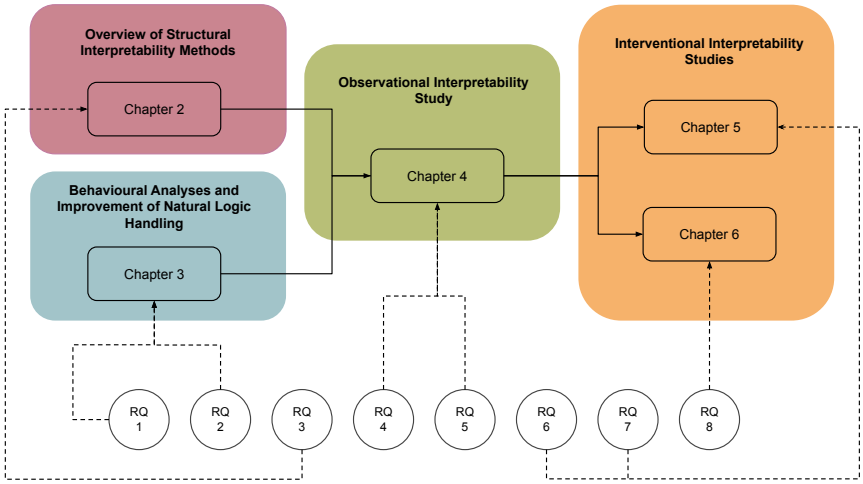


Figure 1.2: Overall structure of the thesis and dependencies between the chapters, as well as connections to the numbered research questions.

C 7. A study of direct and indirect causal effects related to context monotonicity and concept inclusion relations, following the framework of Stolfo, Jin, Shridhar, *et al.* [2] in using these measurements as indicators of *robustness* and *sensitivity*.

1.5 Thesis Outline

The thesis is organised as follows:

Chapter 2 summarises existing work on model interpretability and weighs up the usefulness of previous approaches to our problem setting. The aim of this chaper is to provide an overview of the greater interpretability discussion within which we operate, and motivate the decisions we have made in choosing the experiments carried out throughout this work.

Chapter 3 contextualises many of our ensuing experiments with a summary of existing work specific to the behavioural evaluation of natural logic phenomena in NLI, highlighting observations that context monotonicty is poorly handled by contemporary models. In this chapter, we also present an intermediate context monotonicity detection task which is posited as a potential strategy for supporting better modelling of context monotonicity.

Chapter 4 introduces the NLI-*XY* dataset, following on from the described structure of the natural logic problems of interest. This dataset serves as the basis for a suite

of structural interpretability experiments and qualitative analyses, with a focus on comparing state-of-the-art NLI models to versions which are improved in their natural logic handling capabilities.

Chapter 5 builds on the contributions in chapter 4 by extending the probing studies to *interventional* probing studies, applying an existing method and introducing a new variation better suited to our setting, in light of new observed limitations of the existing approaches.

Chapter 6 takes an alternative approach to the interventional probing methods in chapter 5 and instead views the problem through a causal lens. Specifically, the chapter includes the presentation of a causal diagram which lets us apply frameworks that interpret certain causal effect measures as indicators of robustness and sensitivity to our intermediate semantic features.

Lastly, **Chapter 7** ties together our findings and discusses key limitations which suggest avenues of future work.

1.6 Publications

- **Chapter 3: Supporting Context Monotonicity Abstraction in Neural NLI Models**.
  Julia Rozanova, Doborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, André Freitas. *NaLOMa workshop at IWCS 2022*

- **Chapter 4: Decomposing Natural Logic Inferences in Neural NLI**.
  Julia Rozanova, Doborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, André Freitas. *BlackBoxNLP at EMNLP 2022*

- **Chapter 5: Interventional Probing in High Dimensions: An NLI Case Study**.
  Julia Rozanova, Marco Valentino, Lucas Cordeiro, André Freitas. *Findings of EACL 2023*

- **Chapter 6: Estimating the Causal Effects of Natural Logic Features in Neural NLI Models**.
  Julia Rozanova, Marco Valentino, André Freitas.

Chapter 2

Structural Interpretability for NLP

This chapter addresses **RQ 2** (*Which interpretability methods are best-suited for our interest in detecting emergent intermediate features?*) by presenting a synthesis of previous structural interpretability studies of linguistic features in NLP models (specifically focusing on *probing*), as well as a summary of recent methodological discussions on how best to perform probing studies in a way that decouples the contribution of the probing task and probe itself from the already emergent information in the studied models' representations. Lastly, it includes an overview of how interpretability methods have been extended to *interventional* approaches.

2.1 Introduction

The aim of post-hoc model interpretability work centers around the following form of questions:
Given a neural model trained for an NLP task,

- Which aspects of the end task is the model more or less capable of performing?
- What information has the model learned to extract from its inputs?
- How and where does this information become apparent in the model?
- How does the learned information (such as distinguishable properties and features) inform the model predictions?

We distinguish between *behavioural* and *structural* approaches to answering the above questions. Behavioural studies aim to draw conclusions from looking only at

characteristics of the textual inputs compared to various output measures related to the given end task. Normally, this takes the form of targeted evaluation sets: specifically, task "fragments" that aim either to expose hypothesized heuristic exploits [7], [31] or to isolate a testable aspect of idealized behaviour [12], [13], [36].

On the other hand, structural interpretability emphasizes the study of the intermediate functions and representations within the model. Improved capabilities of NLP models have resulted in increased interest in more in-depth analyses as to the intermediate features and reasoning patterns learned by models. Some such strategies have been quite structure-specific, such as visualisation of neuron activations [37]–[39], examining attention patterns [40], and gradient-based approaches such as saliency analysis [41]. More recently, much focus has been placed on directly examining the *intermediate representations* of a given set of inputs, which is a more architecture-agnostic approach (given that intermediate hidden vector states is a shared structure between many architectural setups). This is done either through projection and visualisation or (more quantitatively) through *probing* [32]. None of the above have been without critical scrutiny, with various works questioning their usefulness, faithfulness and manipulability [32], [42]–[46].

In particular, we have found great utility and promise in the recently burgeoning area of *probing* language model representations. Given early representation visualisation observations in our problem space, probing has been a natural fit and a point of entry into an active discussion on interpretability methodology.

As such, the ensuing in-depth overview in section 2.2 focuses almost entirely on the goals, applications and methodological trends in probing that have informed large parts of the work in this thesis. For a broader overview of interpretability methods in NLP, we recommend Belinkov and Glass [16], Madsen, Reddy, and Chandar [47] or Belinkov, Gehrmann, and Pavlick [48].

Finally, we make one more distinction which captures an overarching trend in both this work and the greater interpretability landscape: the shift from *observational* to *interventional* interpretability studies. We refer to interpretability methods which focus on the post-hoc examination of static values (representations, attention weights, etc) as *observational* studies (also referred to as "inspection-based" studies in Tenney, Das, and Pavlick [10]). These allow for identifying correlational patterns and making interpretability hypotheses, but as many works have pointed out, they are limited in their capacity to support any causal claims; there is no guarantee that observed presence of features are in fact *used* by the model for a given end-task, and in fact the opposite

has been observed [49]. This shortcoming ushers in the later wave of *interventional* interpretability methods, which we will turn to in section 2.3. In contrast to observational studies, interventional studies introduce a modification to either the raw model input, the encoded input representation or the model's strucuture (e.g. weight values) in order to observe the effects of certain changes on model predictions.

2.2 Observational Methods: Representation Probing

At its core, the *probing* methodology can be described as "associat[ing] internal representations with external properties, by training a classifier on said representations that predicts a given property" [32]. It is a strategy that allows us some insights into the abstractions internally encoded by neural models, and in so doing to demonstrate "qualitative properties of the learned representations" [50].

The word "probing" has at times been used in a wider sense than we apply here: it has been used to refer both to behavioural strategies like *diagnostic test sets* and to smaller-scale secondary reasoning task which do not feature any additional training, such as in Richardson, Hu, Moss, *et al.* [12] or Talmor, Elazar, Goldberg, *et al.* [51]. Throughout this work, we use "probing" to mean the training of external machine learning models which take intermediate representations or sets of model weights as their inputs, in order to investigate their structure.

The choice of representations probed depends on the model task and its architecture; earlier studies probe *sentence-level* representations (as in [52]), while the advent of contextualised word vectors from sequence-based models such as BiLSTMs [53] and Transformers [3] has ushered in a greater interest in probing word-level (usually, pooled token-level) representations (as in [54], [55]).

A more exhaustive and general survey of the application and criticisms of probing classifiers can be found in Belinkov [32], and while we cover a similar group of work, we aim to isolate trends and discussion points which help us answer two questions relevant to our research objectives:

- Is the natural logic setting bringing any novelty or advantages to the landscape of structural interpretability studies in NLP?
- Which methodological approaches and perspectives are best suited to design a probing study for our natural logic setting?

We proceed in two directions. Firstly, in section 2.2.1 we glance at the nature of probing tasks that have been of interest in probing studies from the time period leading up to the start of the research presented in this thesis, and position our area of interest with respect to these. In particular, we highlight a predisposition for probing task-agnostic linguistic features, while we are interested in NLI task-specific features.

Secondly, in section 2.2.2 we discuss the ways in which probing methodologies have varied, and address the widely-discussed nuances in the goals of probing studies that have informed the practices we choose to adapt in our own work.

2.2.1 Probing Tasks

In this section, we sketch the trends in the types of model objectives examined vs the intermediate feature identification tasks being probed for.

We call the final task for which the model in question has been trained on the *model objective*, although we acknowedge that (in line with the general state-of-the-art for NLP tasks) most models have multiple training objectives in their lifespan; for example, roberta-large-mnli is really *pretrained* on a language modelling objective before being further *fine-tuned* on the NLI objective, but we refer to the final fine-tuning task as the model objective.

By *probing task* we refer to an auxiliary task which aims to detect intermediate lower-level features which may (directly or indirectly) be useful for the model objective, or at least arising from one of the model's previous training objectives. The structure of probing tasks varies as much as that of legacy NLP tasks. Much of the later methodological discourse centers *classification tasks* for probing, but many traditional NLP tasks (along with their associated metrics, and even custom prediction architectures [55]) have been reformulated as probing tasks; as such, probing tasks can extend beyond classification to (for example) tree-structured prediction tasks [11].

We distinguish between superficial tasks (■), basic syntactic tasks (●), compositionally/hierarchically structured syntactic tasks (◆) and semantic tasks (★). As mentioned in [52], we observe that these separations are somewhat arbitrary, but attempt to be as consistent as possible with the task categorisations in the source material [1].

At the most superficial level, models have been probed for features such as sentence length ($■_1$), word order ($■_2$) and lexical identity ($■_3$) [52], [57]. A large majority of

[1] Some categorisations have contradicted each other: in Conneau, Kruszewski, Lample, *et al.* [52], number prediction (classifying as singular/plural) is considered a *semantic* task, while Linzen, Dupoux, and Goldberg [56] treat it as a syntactic one.

Probed Feature	Reference Symbol	Probed Feature	Reference Symbol
Superficial		**Comp/Hierarch**	
Sentence Length	■1	Span Constituent Labels	◇1
Word Order	■2	CCG Supertagging	◇2
Lexical Identity	■3	Syntactic Ancestor	◇3
		Tree Depth	◇4
		Constituency Parse Tree	◇5
		Arithmetic Intermediate Values	◇6
Syntactic		**Semantic**	
Part of Speech	●1	Coreference	★1
Syntactic Chunking	●2	Semantic Roles / Semantic Proto-Roles	★2
Grammaticality	●3	Named Entities	★3
Conjunct Identification	●4	Dependencies	★4
Bigram Shift	●5	Similarity/Analogy	★5
Number	●6	Semantic Odd-Man-Out	★6
Tense	●7	Coordinate Inverstion	★7
		Concept Relations	★8
		World Knowledge	★9

Table 2.1: Reference table for shorthand symbols to identify the linguistic *features* probed in various probing studies summarised in table 2.2.

probing studies include *syntactic* tasks: POS tagging (●1) is strongly represented in probing studies ([10], [11], [52], [54], [58] and many more), while others include tasks such as syntactic chunking (●2), grammatical error detection (●3) and conjunct identification (●4) [58].

There have also been some syntactic tasks that touch on elements of hierarchical/-compositional syntactic structure without quite reaching the shape of a full parsing task: this includes identifying span/high level constituent labels (◇1) [10], [50], [52], [54], CCG supertagging (◇2) [58], syntactic ancestor prediction (◇3) [58], and tree depth (◇4) [11], [52]. The question of whether full constituency parse trees can be recovered from contextual representations (◇5) has been explored, but requires some additional sophistication in probe structure [11], [50]. In Hewitt and Manning [11], the authors train linear probes (emphasizing low-rank linear transformations) with a training objective that expects distances between word representations to be predictive of parse tree edge numbers.

A greater interest in semantic properties came a bit later; this was in part due to a shift from sentence representations to word-level representations allowing for for "edge"-style probing tasks [54], where pairs of word-level representations were

	Probed Features				Model Objective							
	Superficial	Syntactic	Comp/Hierarch	Semantic	MT	Autoencoder	LM/MLM	Seq2Tree	NER	SRL	Arithmetic	NLI
[56]Linzen et al, 2016		●3, ●6					x					
[59]Shi et al, 2016		●7	◆1		x							
[57]Adi et al, 2017	■1, ■2, ■3					x						
[60]Belinkov et al, 2018		●1		★3	x							
[61]Giulianelli et al, 2018			◆6								x	
[52]Conneau et al, 2018	■1, ■3	●1, ●5, ●6, ●7	◆1, ◆4	★6	x	x		x				x
[50]Peters et al, 2018		●1	◆5	★1			x	x	x	x		x
[62] Zhang et al, 2018		●1	◆2		x		x					
[54]Tenney et al, 2019 (a)		●1,	◆1	★1, ★2, ★3, ★4,★8			x					
[10]Tenney et al, 2019 (b)		●1,	◆1	★1, ★2, ★3, ★4,★8			x					
[58]Liu et al, 2019		●1, ●2, ●3, ●4	◆2, ◆3	★1, ★2, ★3, ★4			x					
[11]Hewitt et al 2019 (a)		●1					x					
[63]Jawahar et al 2019	■1, ■3	●1, ●5, ●6, ●7	◆1, ◆4	★6			x					
[64]Hewitt et al 2019 (b)			◆5				x					

Table 2.2: An overview of *linguistic features* probed for in early probing studies and the *model objectives* of the examined models.

probed for information about pairwise relations, such as coreference (★1) [50], [54], dependencies (★4) [54], semantic roles and similar semantic tags (★2) [54], [58], lexical relations [55] and higher-level entity relations, such as "cause-effect" relations [54] (we group these and other similar tasks under "concept relations" ★8). Non arc-structured semantic word-level/span-level properties probed in a similar vein included named entites (★3) and word similarity (★5).

Task-Specific vs Task-Agnostic Features Earlier probing and interpretability work (loosely speaking, pre-BERT) which focused on sentence representations studied a diverse variety of base model objectives, including models with NLI, machine translation and autoencoder objectives. Ensuing works focus more on models trained on the *masked language modelling* objective (in the case of BERT models, this is is alongside a next sentence prediction objective). This is a deviation from the more colourful pre-BERT landscape, and we are eager to advocate a return to greater diversity in studied model objectives.

In particular, this is because one of the greater aims of interpretability work is to feed into arguments about model *reasoning* strategies: however, the early emphasis on probing *task-agnostic* features leaves little room for speculation about *how* these features are used to inform the model predictions. As language modelling is such a complex and multi-faceted task, it is especially limited in this sense.

In general, it would be useful to choose model objective/probing task pairs which have a clearer (expected) relationship: if we can investigate task-specific features in task-specific models, with prescribed ideas of how the features should be impacting model outputs, the road is clear for more testable interpretability and reasoning hypotheses. We find natural logic to be a suitable setting for such a study, especially because it

finds a balance between naturalistic NLI examples which rely on a significant amount of linguistic complexity (via context monotonicity and noun phrase modification) and world knowledge (via lexical relations), while still boiling down to a reasoning paradigm which can be represented with a simple causal diagram.

Despite the plethora of linguistic feature probing works, not many of these follow this kind of structure. A few examples that are the closest in spirit include Giulianelli, Harding, Mohnert, *et al.* [61], which probes intermediate calculation values of an LSTM-based arithmetic model. In particular, they test hypotheses about the *order* in which intermediate values are calculated. We mention two more, which have also noticed the useful structure of natural logic tasks:

Existing/Contemporary Work On NLI and Natural Logic In strongly aligned contemporary work in Geiger, Richardson, and Potts [15], the authors also apply interpretability approaches to their own natural logic subtask. However, their structural probing applies only to a lexical relation label (also not accounting for phrase-level concept relations), and not the monotonicity profile. While their dataset does include a distinction between upward and downward monotone settings for their lexical substitutions, the only downward monotone operator present is the negation operator "not", while we incorporate a greater variety of downward monotone operators. More recently, Jumelet, Denic, Szymanik, *et al.* [65] did some probing experiments on detecting polarity of monotonicity environments. However, they do not study entailment/NLI models, focusing rather on a language modelling objective, and are interested in the impact of monotonicity on *negative polarity item licensing*.

2.2.2 Methodological Trends and Developments

Early work on probing was naturally much influenced by the pre-2019 NLP paradigm, where independently trained word representations (e.g. word2vec [66], GloVe [67], ELMO [68]) were used as inputs to machine learning models for downstream NLP tasks. Embedding "quality" was tied to how well the representations allowed for potential secondary model performance maximisation on a downstream task. The structure is in many ways identical to probing: static representations (trained using a different objectives and architectures) are used to train a new model for a different NLP task. It is expected that similar experimental setups were initially used for probing, such as the comparison to human performance upper bounds [52] and even using the same architectures for probes as for the original downstream NLP tasks they were designed

for [55].

The greater theme of this section is to demonstrate how the field identified the subtly different goals of probing, and moved away from older experimental setups to new constraints and best practices. The key question that has begun to dominate probing studies is to what extent intermediate features are *emergent* and thus *easily extractable* in learned representations, rather than how much *capacity* the representations have to be predictive of a secondary feature given *an arbitrary amount of additional training*.

Throughout this growth process, experimental setup for probing studies has been broadly scattered, with design choices reflecting slight differences in goals and perspectives. The streamlining of methodological standards for probing has been an ongoing topic of discussion and reasearch throughout the years spanning this thesis project. We describe the variations, insights and suggestions that have influenced our work.

We consider four dimensions of variation: baselines and control tasks, layer-wise dynamics, probe complexity and training dynamics. We present a timeline of the mentioned works in table 2.3.

1. **Baselines and Controls**

   Efforts to contextualise probing results by comparing their capabilities to reasonable baselines have been present throughout most probing works, but the choice of baselines has not been in any way homogenous across studies. Baseline choices have included either lower-bound representational baselines which are known to capture specific aspects of relevant information (such as a TF-IDF-based representations [52]) or randomised representations [52], [54], [62]. Upper bounds are less prevalent in probing, but two examples include the use of a human task performance upper bound [52] or a higher-complexity probe constructed in a similar way to a prior state-of-the-art model for a specific task [58]. Choice of baselines is also narrative-dependent: for example, studies wishing to isolate contextual influence (e.g. [54], [55], and [58]) included representational baselines fron non-contextual word encoders such as GloVe ([67]).

   Closely linked to the practise of a *representational baseline* (known to capture a minimal amount of information that allows for a simplistic approach to the task) is the use of *control tasks*. Hewitt and Liang [64] design a control task for POS tag probing by randomly re-labelling the POS dataset in a structured way that demonstrates the "most frequent tag" (MFT) strategy: namely, assigning the same tags to all shared surface word forms across the train and test set. In a

similar spirit, Pimentel, Valvoda, Hall Maudslay, *et al.* [69] advocate designing a *control function*, but theoretical work in Zhu and Rudzicz [70] argues that this is information theoretically equivalent to the control tasks in Hewitt and Liang [64].

Many proposed probing metrics incorporate the ideas of baselines or controls by subtracting the accuracy of a control task ("selectivity", [64]), subtracting the accuracy of a baseline representation in the form of a previous layer, or a concatenation of the previous layer with the target layer ([71], [72]). We discuss the use of previous layers as baselines more in the context of layer-wise dynamics in the next section.

2. **Layer-Wise Dynamics**

Beyond static auxiliary tasks applied to a fixed choice of representation, there is interest in *where* certain types of information emerges across layers ([10], [60], [71], [72]), and whether this is in line with more traditional NLP pipelines or other expectations for order in linguistic information types [10], [72].

At its simplest, this can be an inclusion of a probing score at each layer [11], [55], [60], In more mature strategies, these lines of investigation can include their own adapted metrics and probe architectures: for example, Tenney, Das, and Pavlick [10] use *scalar mixing weights*, which allow them to simultaneously probe multiple layers with tunable layer weights. These in turn serve as a metric for layer relevance for a given probing task. Layer weights are summarised with a "center of gravity" metric, serving as an expected value for the most impactful layer. They also provide a cumulative version of their score, where the scalar mixing strategy is applied to all layers up to a given layer n.

Other works turn to earlier model layers as a baseline for demonstrating information gain across layers: Hewitt, Ethayarajh, Liang, *et al.* [71] substract probing scores for a given layer from the zero-th layer, while Kunz and Kuhlmann [72] adapt these metrics to condition on the previous layer.

A recurring observation in these order-focused studies is that lower-level, syntactic information is more present in earlier layers, and can even be "forgotten" in later layers. Meanwhile, higher-level semantic information is more prevalent in later layers [10], [63].

Liu, Gardner, Belinkov, *et al.* [58] introduce a notion of comparative *transferability* of layer representations; in their description, layers which achieve greater probing scores across a large variety of tasks are deemed to be "more transferable". They posit that recurrent architectures have early layers demonstrating the most transferability, while transformer models differ more widely, with a greater spread of task competence between layers.

3. **Probe Complexity**

   The size, complexity and architectural shape of probes differ wildly across studies. From low-rank linear classifiers, MLPs + ReLU to full custom neural architectures designed for a specific task [55], it becomes exceedingly difficult to view probing scores in isolation, or to compare them across works.

   In one of the earlier works to include *multiple* probe architectures/complexities, Liu, Gardner, Belinkov, *et al.* [58] noticed that poor probe performance on certain tasks such as NER was strongly boosted by using a MLP model with ReLU activations, as opposed to a simpler linear model with fewer parameters. Even more strikingly, Zhang and Bowman [62] showed that their randomly initialized representations supported comparably high probing accuracies to their interpretability target (with sufficient probe training). In their influential work, Hewitt and Liang [64] verbalize the ambiguity between the extent to which one may attribute probing accuracy to the *input representation* or to the contribution to the trained probe itself.

   This kicked off a discussion in favour of formally defining the more subtle goals of probing: in particular, disentangling the presence of already *emergent* linguistic information in model representations and the *capacity* of the union of representation, probe and probing task data to achieve a high probing accuracy score. Suggestions have included the introduction of new metrics, reporting results for a variety of probe complexities, using the simplest possible probes, or even using the most complex available probes [69].

   Theoretical discussions regarding the role of *information-theoretic* measures have dominated this discourse as a framework for supporting advances in probing methodologies. In Voita and Titov [73], the authors choose to operationalize the notion of *ease of extractability* as their quantity of interest, introducing an information-theoretic *minimum description length* measure. A common theme has been to use the *mutual information* between the representations and probing

task labels as a standard quantity to estimate, guiding works such as Pimentel, Valvoda, Hall Maudslay, *et al.* [69] and Pimentel, Saphra, Williams, *et al.* [74] and Zhu and Rudzicz [70]. However, this has not gone unquestioned, and alternatives such as *Bayesian mutual information* have been proposed [75].

4. **Probe Training Dynamics** A practical way to adress the concerns of probe complexity has been to examine various aspects of probe training dynamics. A common suggestion has been to use much smaller proportions of the probing task data for training than for testing, as in Kunz and Kuhlmann [76]. An even more rigid variation on this theme is to log accuracy at intermediate epochs [52] of the training process and compare the *speed of learning* evident in the curves (again connecting to the notion of "ease of extractability".) Indeed, one of the minimum description length metrics in Voita and Titov [73] requires gradually incrementing the size of the training set.

Emergence vs Capacity In summary, rather than merely measuring the joint capacity of representations and probing models to perform well at a linguistic task, probing studies should aim to build arguments for the already complete *emergence* of linguistic information in model representations from the original model training objective. There is still no particular consensus on best practices or universal adoption of metrics, but many of the varied suggestions have added to a shared toolbox for building a convincing interpretability argument. However, most works agree that claims should be constrained to comparative assessments across applications of the same probing methodologies.

2.3 Interventional Methods

Even when strong arguments are built for the presence of certain types of linguistic information within model representations, this does not necessarily constitute evidence that the identified information is in fact used by the model in its latent decision-making process. In fact, observations in Ravichander, Belinkov, and Hovy [33] note the opposite, identifying the presence of features that serve no use for a given end-task.

In summary, the above observational methods are correlational in nature [16], while claims of feature use for model predictions are strictly causal claims. In line with ideas of causal modelling [77], it is only through *interventional* studies that we can build arguments for causal effects. In particular, we are especially concerned with abstract

Date (MM/YY)	Baselines and Controls	Layer-Wise Dynamics	Probe Complexity	Probe Training Dynamics
2018				
09/18	Conneau, Kruszewski, Lample, *et al.* [52]			Conneau, Kruszewski, Lample, *et al.* [52]
11/18	Zhang and Bowman [62]			
2019				
05/19	Tenney, Xia, Chen, *et al.* [54]	Tenney, Das, and Pavlick [10]		
06/19	Liu, Gardner, Belinkov, *et al.* [58]	Liu, Gardner, Belinkov, *et al.* [58], Hewitt and Manning [11]	Liu, Gardner, Belinkov, *et al.* [58], Hewitt and Manning [11]	
07/19			Jawahar, Sagot, and Seddah [63]	
09/19	Hewitt and Liang [64]		Hewitt and Liang [64]	
2020				
07/20	Pimentel, Valvoda, Hall Maudslay, *et al.* [69]		Pimentel, Valvoda, Hall Maudslay, *et al.* [69]	
08/20	Vulić, Ponti, Litschko, *et al.* [55]	Vulić, Ponti, Litschko, *et al.* [55]		
11/20				
2021				
01/21			Voita and Titov [73], Pimentel, Saphra, Williams, *et al.* [74], Zhu and Rudzicz [70]	Voita and Titov [73]
11/21	Hewitt, Ethayarajh, Liang, *et al.* [71]	Hewitt, Ethayarajh, Liang, *et al.* [71]	Pimentel and Cotterell [75]	Kunz and Kuhlmann [76]

Table 2.3: A timeline-structured schematic listing the works mentioned along each dimension of probing methodology variations. As far as possible, we use the date associated with the corresponding ACL anthology entry.

semantic and linguistic features: if we have already shown (using structural interpretability methods) that certain concepts are captured by the model, what interventions can we perform to demonstrate feature influence? The structure of studies which aim to answer this question follow a general setup of intervention and outcome measurement, where the intervention may take place at either the input or representation level.

An early use of an interventional strategy in model interpretability is Giulianelli, Harding, Mohnert, *et al.* [61], where the probing results of subject-verb agreement are compared before and after gradient-driven representation modifications.

A key work following up on probing methodology is the *interventional probing* approach of Elazar, Ravfogel, Jacovi, *et al.* [1], which followed the methodology of Ravfogel, Elazar, Gonen, *et al.* [78] to eliminate detected feature information by the process of *iterative nullspace projection*. The measured outcome is the final model task performance. However, various critical studies have pointed out limitations in this work, such as Kumar, Tan, and Sharma [79] and our own work in chapter 5, which introduces a related interventional variation which we call *mnestic* probing. A relevant work which also focuses on NLI is Geiger, Richardson, and Potts [15], which falls somewhere between a structural and a textual input-level intervention: they propose *interchange interventions*, where they replace tokens in the representation sequence with the values they would have taken under a different input.

Not all interventional interpretability methods deal directly with causal measures and causal diagrams, but there has been a strong trend in this direction: strong arguments for model use of gendered features are supported by a rigid *causal mediation analysis* in Vig, Gehrmann, Belinkov, *et al.* [34] and Finlayson, Mueller, Gehrmann, *et al.* [35], where causal effect measures are used to estimate the extent to which given features are mediated through specific architectural components.

While the calculation of the causal treatment effect measures of interest is not always feasible, there have been informative studies on the causal effects that can be calculated: for example, Amini, Pimentel, Meister, *et al.* [80] use naturalistic input interventions to measure the effects on the representations themselves, while Stolfo, Jin, Shridhar, *et al.* [2] investigate unwanted direct effects as proxy measures for robustness and sensitivity.

Beyond merely testing whether identified embedded concepts have a causal influence on the model, works such as Geiger, Lu, Icard, *et al.* [81] draw on *causal abstraction theory* and search for an alignment between model weights/representations and easier-to-understand, conceptual causal models which are as *faithful* as possible to the model's actual behaviour (see also Jacovi and Goldberg [46] and Geiger, Wu, Lu, *et al.* [82])

2.4 Conclusion and Influence on Our Work

The observations in this section address **RQ 2** (*Which interpretability methods are best-suited for our interest in detecting emergent intermediate features?*), and motivates the choice of general research directions and specific interpretability strategies we follow and build on throughout this work.

Firstly, The lack of granular investigations into semantic features specific to the

NLI task has encouraged us to dig deeper into emergence of natural logic features in transformer-based NLI models. In constructing our probing study, as we wish to emphasise the clear *emergence* of context monotonicity features after the HELP fine-tuning strategy and demonstrate that this feature becomes comparatively easy to extract, we have most taken to heart the practise of varying probe complexity (adopting the nuclear norm approach in Pimentel, Saphra, Williams, *et al.* [74]) and comparing probing accuracy curves across models. We also adopt the selectivity metric in Hewitt and Liang [64].

Secondly, we follow the call for a greater emphasis on *interventional* methods in order to support stronger claims about model reasoning strategies. In particular, our two papers on interventional methods are most influenced by Elazar, Ravfogel, Jacovi, *et al.* [1] and Stolfo, Jin, Shridhar, *et al.* [2] respectively.

2.5 Scoping and Limitations

The scope of section 2.2.1 is limited to the state of affairs in the first year of this PhD project (2019) and the most relevant previous work related to neural NLP models and their structural interpretability. This reflects the initial assessment as to whether a larger body of structural interpretability work in the NLI space would fill a relevant gap, although we add a note on some contemporary works that have arisen in the same space. Sections 2.2.2 and 2.3 cover works roughly throughout the duration of this project (2019—2022), reflecting the evolving views and methodologies in structural interpretability that we have dialogued with in our work. It must be noted that many interpretability methods have their roots in work on computer vision, but we limit our collected works here to works in NLP.

Chapter 3

Supporting Context Monotonicity Abstraction

In this chapter, we address our first empirical research question, **RQ 1**: *How much does fine-tuning on the HELP dataset improve NLI models' performance on existing natural logic evaluation datasets? Would a secondary transfer-learning task based on the prediction of context monotonicity result in an improvement in overall evaluation scores?*.

We begin with a summary of existing behavioural evaluation of natural logic phenomena in NLI, highlighting observations that examples requiring context monotonicity judgments are poorly handled by contemporary models, suggesting that it is an absent feature until improvement strategies (such as fine-tuning on the HELP [14] dataset) are introduced. Next, we introduce introduce a potential complementary improvement strategy: a context monotonicity predicition task as a secondary training objective in the style of *transfer learning*.

We perform a set of experiments which allows for the comparison of these improvement strategies, examining performance on existing evaluation sets. We find that introducing a transfer learning step with our context monotonicity prediction task does not significantly boost performance further than the improvements already introduced by fine-tuning on the HELP dataset.

3.1 Introduction

NLI has seen much success in terms of performance on large benchmark datasets, but there are still expected systematic reasoning patterns that we fail to observe in

the state-of-the-art NLI models. We focus in particular on the class of NLI problems defined in section 1.2.1, which can be described as a form of *substitutional* reasoning which displays logical regularities with respect to substitution of related concepts. This reasoning pattern (referred to as *monotonicity reasoning* in relevant works such as Yanaka, Mineshima, Bekki, *et al.* [13]) is systematic, and thus is a much-tested behaviour in enquiries into the *systematicity* and *generalisation capability* of neural NLI models [12]–[14], [83], [84].

Monotonicity reasoning has a history of causing problems for neural NLI models: it has been observed [14], [84] that current state-of-the-art transformer-based NLI models tend to routinely fail in downward monotone contexts, such as those arising in the presence of negation or generalised quantifiers. This suggests that models are failing to capture the effect of context monotonicity on the entailment label. Recent strategies [12] to address the shortcomings of NLI models in downward-monotone contexts have followed the *inoculation* method [85]: additional NLI training data which provides examples of the target phenomenon (in this case, downward-monotone reasoning) is used to fine-tune existing models. This is done with some success in Richardson, Hu, Moss, *et al.* [12] and Yanaka, Mineshima, Bekki, *et al.* [14] and [84].

In contrast, we wish to investigate a *transfer learning* strategy that directly targets monotonicity classification as an additional training task (see the schamatic in figure 3.1) to see if this can further improve the monotonicity reasoning performance of popular transformer-based NLI models (when the model is fine-tuned again for the NLI objective).

Our contributions are as follows:

- Extending our description of contexts as abstract units in section 1.2.1 to an experimental setting, we introduce an improvement in neural NLI model performance on monotonicity reasoning challenge datasets by employing a context monotonicity classification task in the training pipieline of NLI models. To the best of our knowledge, this is the first use of neural models for this specific task.

- For this purpose, we adapt the HELP dataset [14] into a HELP-Contexts dataset, isolating contexts and their monotonicity labels.

- For the class of NLI problems described as *monotonicity reasoning*, we demonstrate the impact of the proposed transfer strategy: we show that there can be a slight improvement on downward monotone contexts (on top of existing improvement strategies), previously known to be a bottleneck for neural NLI models. As

such, this shows the possible benefit of directly targeting intermediate abstractions (in this case, monotonicity) on which the final label depends.

3.2 Related Work

The study of monotonicity in natural language has a strongly developed linguistic and mathematical theoretical groundwork, dating back to the monotonicity calculus of Sanchez [21] and in semantic studies such as Van Benthem *et al.* [86]. The inferential mechanism based on monotonicity properties of quantifiers, determiners and contexts in general is referred to either as *natural logic* or *monotonicity reasoning*.

There are varying formal and informal presentations and some variation in terminology, but the format most relevent to our work is the presentation of the output out of the *ccg2mono* system by Hu and Moss [24] (the monotonicity tagging system on which the *HELP* dataset has relied on for its construction). In their presentation, the concept of monotonicity is represented in terms of *polarity tags* assigned to words in a sentence, as per their example:

(5) Every dog$^{\downarrow}$ scares$^{\uparrow}$ at least two$^{\downarrow}$ cats$^{\uparrow}$.

Any word labelled with an $\uparrow$ tag can be replaced with a more general word (likewise, with a more specific word when labelled with a $\downarrow$ tag), yielding a sentence which is entailed by the starting sentence. Note that the polarity of a word is equivalent to our definition of the monotonicity of its context in section 1.2.1: we see this as a more useful perspective, as the polarity tag is determined by the words in the context rather than being a property of the word itself. In practice, their system takes a CCG parse tree of a sentence as its input and assigns polarity tags to words in the sentence by taking into account annotations given to linguistic operators into whose scope the word falls. We refer to Hu and Moss [24] for the full formal and technical details of their polarity tagging system. In our experimental work and the construction of our dataset, we take for granted on the existence of polarity tags which have already been assigned, without needing to delve further into the thread of information that determines the tag.

We now provide some background on approaches which use the ideas of natural logic to either create symbolic inference systems based on formal descriptions of monotonicity phenomena, or to test the handling of natural logic reasoning by trained NLI models.

Symbolic Implementations There are two flavours of implementations that result in the deductions allowed by monotonicity reasoning. Firstly, works such as Hu, Chen, Richardson, *et al.* [87] and Abzianidze [88] rely on linguistically-informed polarity markings on the nodes of CCG parse trees. They require accurate parses and expertly hand-crafted linguistic rules to mark the nodes with polarity tags, as in Hu and Moss [24]. In Hu, Chen, Richardson, *et al.* [87], a premise is tagged for monotonicity and a knowledge base of hypotheses created by a substitution known to be truth-preserving is generated. Candidate hypotheses are compared with this set, checking for exact matches. On the other hand, [88] uses the CCG parses to further translate sentences to a lambda logical form for use in a deduction method inspired by tableau calculus. These approaches differ from strategies such as in MacCartney and Manning [89], which require an *edit sequence* which transforms the premise into the hypothesis. Atomic edits are tagged with generalised entailment relations which are combined with a join operator based on relational composition to determine whether the transformation is overall truth-preserving, hence yielding a hypothesis entailed by the premise. Later, Angeli and Manning [90] treated these atomic edits as edges in a graph and phrased entailment detection as a graph search problem. Concepts from symbolic approaches to NLI have also been applied in symbolic question answering systems (such as in Bobrow, Cheslow, Condoravdi, *et al.* [91]), and hybridized with neural systems (such as in Kalouli, Crouch, and Paiva [92]).

Neural NLI Models and Monotonicity State-of-the-art NLI models have previously been shown [14], [84] to perform poorly on examples where the context f is *downward monotone*, as occurs in the presence of negation and various generalised quantifiers such as “every” and “neither” (with more examples of downward monotone operators in table 1.3). Benchmark datasets such as MNLI are somewhat starved of such examples, as observed by Yanaka, Mineshima, Bekki, *et al.* [14]. As a consequence, the models trained on such benchmark datasets as MNLI not only fail in downward monotone contexts, but *systematically* fail: they tend to treat all examples as if the contexts are upward monotone, predicting the *opposite* entailment label with high accuracy [14], [84]. A few datasets have been introduced to test the performance of both neural and symbolic systems on sets of natural logic examples: we tabulate their use across the abovementioned works in table 3.1.

Data augmentation techniques and additional fine-tuning with an inoculation [85] strategy have been attempted in Richardson, Hu, Moss, *et al.* [12], Yanaka, Mineshima,

Bekki, *et al.* [14], and Geiger, Richardson, and Potts [84]. In the latter case, performance on a challenge test set improved without much performance loss on the original benchmark evaluation set (SNLI), but in Yanaka, Mineshima, Bekki, *et al.* [14] there was a significant decrease in performance on the MNLI evaluation set. These studies form the basis on which we aim to build, and their choice of evaluation datasets and models inspires our own choices.

		Previous Work			
Evaluation Datasets		Geiger 2020 (Neural)	Yanaka 2020 (Neural)	Moss 2019 (Neural)	Hu 2020 (Symbolic)
Large, Broad Coverage	MNLI Test		x		
	MNLI Dev (Mismatched)			x	
	SNLI Test	x		x	
Small, Targeted Phenomena	MED		x		
	SICK		x*		x
	FraCaS		x*		x
	MoNLI Test	x			
	Monotonicity Fragments			x	x

Table 3.1: Evaluation datasets used in previous work investigating monotonicity reasoning. Positions marked $*$ indicate that the dataset is included in another used evaluation dataset.

Neural Transformer-based language models have been shown to implicitly model syntactic structure [93]. There is also evidence to suggest that these NLI models are at least representing the concept relations quite well and using this information to predict the entailment label, as corroborated by a study based on *interchange interventions* in [84].

We hypothesise that such models have the capacity for learning monotonicity features. The extent to which the representations capture monotonicity information in the contextual representations of tokens in the sequence is not yet well understood, and this is an investigation we wish to initiate and encourage with this work.

3.3 Experiments

Building on the observations in the above-mentioned previous papers, we ask the following questions:

- Can a context monotonicity classification task in the model training pipeline further improve performance on targeted evaluation sets which test monotonicity reasoning?

- Does this mitigate the decrease in performance on benchmark NLI datasets?

Our investigation proceeds in three parts: Firstly, we attempt to fine-tune a SOTA NLI model for a context monotonicity classification task.

Secondly, we retrain the above model for NLI and evaluate the performance on several evaluation datasets which specifically target examples of both upward and downward monotonicity reasoning. We examine whether there is any improvement over a previously suggested approach on fine-tuning on a large, automatically generated dataset (HELP) from Yanaka, Mineshima, Bekki, *et al.* [14].

Models We start with existing NLI models pretrained on benchmark NLI datasets. In particular (and for best comparison with related studies) we use RoBERTa [5] pretrained on MNLI [18] and BERT [17] pretrained on SNLI [19]. These are two benchmark NLI datasets which contain examples derived from naturally occurring text and crowd-sourced labels, aiming for scale and broad coverage. We do not deviate from the architecture, as we are only investigating the effect of training on different tasks (monotonicity classification and NLI) and datasets.

3.3.1 Retraining NLI Models to Classify Context Monotonicity

Symbolic approaches such as Hu and Moss [24] treat monotonicity classification as the task of labeling words in a sentence with either an upward or downward polarity marking, in the manner of a sequence tagging task. Our emphasis of monotonicity as a property of a *context* allows for a different framing of this problem: we consider monotonicity classification as a binary classification task which takes a context as its input, with an explicit indication (with a variable) of the "slot" in the sentence for which we wish to know the polarity. Different positions of the variable in a partial sentence may yield a context with a different monotonicity label; a typical example of this is sentences featuring generalised quantifiers such as "every", which may be monotone up in one argument but monotone down in another. For the class of NLI problems we consider, it is only the monotonicity of the shared context that matters to the entailment label, so we target only the classification of that monotonicity value.

3.3.1.1 Input Representation

The NLI models which we wish to start with are transformer-based models, in line with the current state-of-the-art approaches to NLI. Transformer models represent a sentence as a sequence of tokens: we take a naive approach to representing a context by indicating the variable with an uninformative 'x' token. We refrain from using the mask token to indicate the variable, as the underlying pretrained transformer language models are trained to embed the mask token in such a way as to correspond with high-likelihood insertions in that position, which we would prefer to avoid.

3.3.1.2 Dataset

In order to ensure our monotonicity classification task does not add any unseen data (when compared to only fine-tuning on the HELP dataset) we adapt the HELP dataset for this task. The HELP dataset is originally constructed by tagging sentences from the Parallel Meaning Bank [94] using the *ccg2mono* [24] polarity tagging system and introduce concept substitutions based on WordNet [95] relation annotations.

The original composition of the HELP dataset can be seen in Table 3.2, originally from Yanaka, Mineshima, Bekki, *et al.* [14].

Section	Size	Example
Up	7784	*Tom bought some **Mexican sunflowers** for Mary* ⇒*Tom bought some **flowers** for Mary**
Down	21192	*If there's no **water**, there's no whisky** ⇒*If there's no **facility**, there's no whisky*
Non	1105	*Shakespeare wrote both **tragedy and comedy*** ⇏*Shakespeare wrote both **tragedy and drama***
Conj	6076	*Tom removed his glasses* ⇏*Tom removed his glasses **and rubbed his eyes****
Disj	438	*The trees are barren* ⇒*The trees are barren **or bear only small fruit****

Table 3.2: Dataset details of the HELP dataset, which we draw on for the HELP-Contexts dataset. The relevant portions are the "Up" and "Down" monotonicity reasoning examples.

We manually select a random sample of 2000 examples evenly split across upward monotone and downward monotone categories, and extracted context examples according to the following criteria:

1. Identify premise, hypothesis pairs which are $(f(x), f(y))$ pairs as per our construction, differing *exactly by one noun phrase*. Any sampled examples which are not structured in this way are discarded.

2. Extract the shared context f, replacing the noun phrase with a 'x' symbol.

3. Assign a gold label of *context monotonicity* according to whether the monotonicity reasoning pattern was labelled as "up" or "down".

As such, we extract only the contexts f and the monotonicity label into dataset which we will call "HELP-Contexts", which we split into a train and test set in a 70:30 ratio (featuring a final training set of 686 examples). Examples of this dataset are presented on Table 3.3. [1]

Context	**Context Monotonicity**
There were no x today.	downward monotone
There is no time for x.	downward monotone
Every x laughed.	downward monotone
There is little if any hope for his x .	downward monotone
Some x are allergic to wheat.	upward monotone
Tom is buying some flowers for x.	upward monotone
You can see some wild rabbits in the x.	upward monotone

Table 3.3: Examples from the HELP-Contexts dataset, with respective labels.

3.3.1.3 Results

As presented in Table 3.4, the task of predicting the monotonicity of the contexts in the HELP-Contexts dataset can be solved using fine-tuned transformer models. This suggests a potential path for inducing a bias for context classification in downstream tasks such as NLI, which could benefit from better encoding of context monotonicity.

3.3.2 Improving NLI Performance on Monotonicity Reasoning

3.3.2.1 Training Data

We use the following datasets for training and evaluation respectively: we begin by once again using the HELP dataset [14], which was designed specifically as a balanced

[1] The original HELP dataset also contains a few non-monotone examples: in the current state of this work, these are omitted in favor of a focus on the specific confusion in existing models where downwards monotone contexts are often treated as upwards monotone ones.

Model	Evaluation Data					
	HELP-Contexts Dev			HELP-Contexts Test		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
bert-base	98.74	99.08	98.91	98.00	95.24	96.54
bert-large	98.23	98.88	98.55	97.51	95.70	96.57
roberta-large-mnli	99.62	98.92	99.26	98.73	96.64	97.64
roberta-large	99.81	99.46	99.27	98.99	96.41	97.62
roberta-base	99.81	99.46	99.63	98.10	95.56	96.76
bert-base-uncased-snli	98.88	98.19	98.53	98.92	97.29	98.07

Table 3.4: Performance of state-of-the-art models for the context prediction task. Each model was trained on HELP contexts (training set).

additional training set for the improvement of NLI models with respect to monotonicity reasoning. We create a split of this dataset which is based on the HELP-Contexts dataset by assigning each example either to the train or test set depending on which split its associated context f is in the HELP-Contexts dataset. This is to ensure there is no overlap between the examples' contexts across the three data partitions. Our approach combined this strategy with an additional step based on the context monotonicity task described in section 3.3.1.

3.3.2.2 Training Procedure

We rely on the architecture implementations and pretrained models available with the *transformers* library [96]. As indicated in the schematic in figure 3.1, we start with the NLI models pretrained on benchmark NLI datasets (which we shall henceforth tag as "bert-base-uncased-snli" and "roberta-large-mnli"). We first fine-tune these models for the context monotonicity classification task using the training partition of the HELP-Contexts dataset. We re-use the classification head of the pretrained models for this purpose, but only use two output states for the classification. Lastly, we fine-tune on the HELP dataset, so that the final model is once again an NLI model and thus may be compared to other NLI models, including those fine-tuned only on the HELP dataset (after benchmark NLI pretraining).

3.3.2.3 Evaluation Data

Evaluation datasets are typically small, challenging and categorized by certain target semantic phenomena. Following previous work in this area, we evaluate our approach

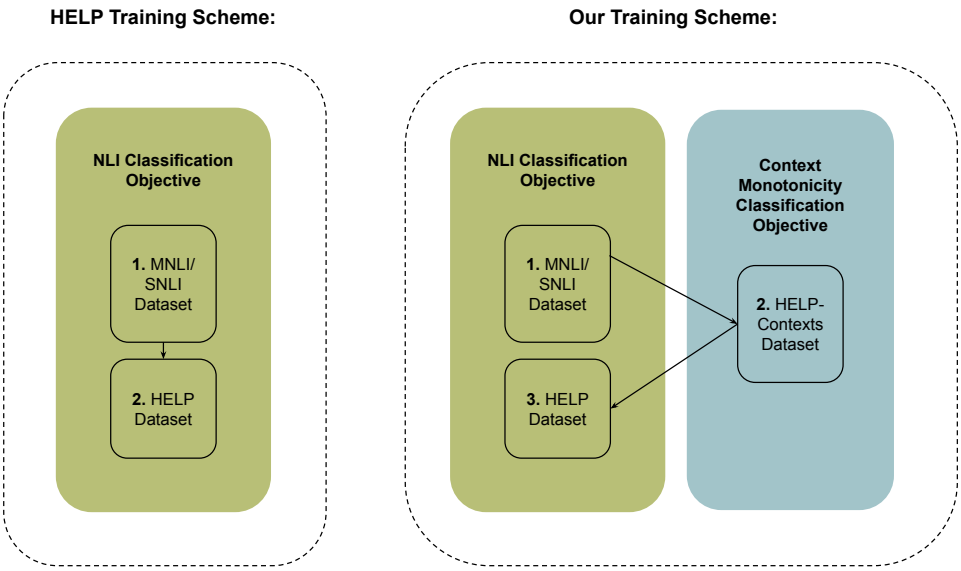


Figure 3.1: Our training scheme introduces a transfer learning step by fine-tuning an NLI model on a secondary context monotonicity classification task, before fine-tuning on the HELP dataset so that we once again end up with an NLI classifier.

using the MED dataset introduced in [13], which is annotated with monotonicity information and draws from various expert-curated diagnostic challenge sets in NLI such as SICK, FraCaS and the SuperGlue Diagnostic set. It features a balanced split between upward and downward monotone contexts, in contrast to the benchmark MNLI dataset. Additionally, we include evaluation on the MoNLI dataset [84] which also features a labelled balance of upward and downward monotone examples. However, the latter dataset's downward monotone examples are only exemplary of contexts featuring the negation operator *"not"*, whereas MED [13] also includes more complex downward monotone operators such as generalised quantifiers and determiners. We refer to these respective papers [13], [84] for full breakdowns and analyses of these datasets.

3.3.2.4 Baselines

Although the main comparison to be made is the improvement introduced when including the context-monotonicity-classification training on top of the current state-of-the-art roberta-large-mnli model trained on HELP, we include additional baselines: roberta-large-mnli fine-tuned on the *monotonicity fragment* from the *semantic fragments* [12] dataset. The strategy in this work is the same as with the HELP dataset, but we include

this in the evaluation on the chosen challenge sets for a more complete comparison.

3.3.2.5 Results

We present the results on the challenge sets MED and MoNLI in Table 3.5, with a break-down by upward and downward monotone contexts. Furthermore, we have re-run each model on the original benchmark evaluation datasets SNLI and MNLI, with the results visible in Table 3.6. We guide through and discuss these results in section 3.4.

Model	**Additional Training Data**	**Challenge Datasets**					
		MoNLI Test			**MED**		
		Upward Mono	Downward Mono	**All**	Upward Mono	Downward Mono	**All**
bert-base-uncased-snli	-	37.74	56.49	46.15	53.58	43.91	49.36
bert-base-uncased-snli	HELP	30.89	85.02	**55.19**	43.4	72.43	60.18
bert-base-uncased-snli	**HELP + HELP-Contexts**	21.6	97.67	**55.19**	32.56	87.13	**66.22**
roberta-large-mnli	-	95.19	5.32	58.84	82.12	25.76	46.09
roberta-large-mnli	Monotonicity Fragments (Easy)	92.68	79.62	86.81	74.54	65.68	70.05
roberta-large-mnli	Monotonicity Fragments (All)	50.00	50.00	50.00	35.42	61.80	49.78
roberta-large-mnli	HELP	94.72	98.67	96.48	64.47	86.25	**77.4**
roberta-large-mnli	**HELP + HELP-Contexts**	98.78	97.17	**98.06**	65.24	85.12	76.44

Table 3.5: Performance of NLI models on challenge datasets designed to test performance on monotonicity reasoning.

3.4 Discussion

Average Performance Firstly, we confirm previous observations that the starting pretrained transformer model roberta-large-mnli (which is considered a high-performing NLI model, achieving over 93% accuracy on the large MNLI development set) has a dramatic performance imbalance with respect to context monotonicity. The fact that performance on downward monotone contexts is as low as 5% suggests that this model perhaps routinely assumes upward monotone contexts. It was noted in Yanaka,

Model	Additional Training Data	Benchmark Datasets							
		MNLI (m*) Dev		MNLI (mm*) Dev		SNLI Dev		SNLI Test	
		Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
bert-base-uncased-snli	-	44.96	-	45.52	-	41.54	-	40.78	-
bert-base-uncased-snli	HELP	35.13	-9.83	34.37	-11.5	25.93	-15.61	25.92	-14.86
bert-base-uncased-snli	**HELP + HELP-Contexts**	36.91	**-8.05**	37.36	**-8.16**	36.54	**-5.00**	37.20	**-3.58**
roberta-large-mnli	-	94.11	-	93.88	-	93.33	-	93.14	-
roberta-large-mnli	HELP	82.66	**-11.45**	83.38	**-10.50**	74.77	-18.56	74.39	-18.75
roberta-large-mnli	**HELP + HELP-Contexts**	81.00	-13.11	82.01	-11.87	82.99	**-10.34**	82.31	**-10.83**

Table 3.6: Fine-tuning state-of-the-art NLI models with the aim of improving monotonicity has tended to result in lower performance on the original benchmark NLI datasets. We compare these performance losses in addition to tracking performance on the the challenge datasets. * MNLI (m) and (mm) refers to the matched and mismatched dataset respectively. For MNLI, only the *Dev* set is publically available.

Mineshima, Bekki, *et al.* [14] that the MNLI benchmark dataset is strongly skewed in favor of upward monotone examples, which may account for this.

We can see in Table 3.5 that our approach (fine-tuning on both HELP-Contexts and HELP) slightly outperforms or matches the baseline models in three of the accuracy scores in the "All" category (achieving 55.19%, 66.22% and 98.06% respectively), and is competitive in the fourth (achieving 76.44%, slightly lower than the highest 77.4% achieved by roberta-large-mnli fine-tuned only on HELP).

Furthermore, in Table 3.6 we observe less performance loss (in most cases) on the benchmark NLI datasets than for the models fine-tuned only on HELP (except in the case of roberta-large-mnli on the MNLI dataset). However, it seems that the majority of gains are already due to the training on the HELP dataset, suggesting that perhaps it is already enough to introduce a useful "understanding" of context monotonicity behaviour.

Performance by Monotonicity Category As evident from Table 3.5, we observe a substantial improvement for the bert-base-uncased NLI models for downward monotone contexts. For the much larger roberta-large-mnli models, any gains over the model trained on HELP only are quite small. A common observation is the notable trade-off between accuracy on upward and downward monotone contexts; training that improves one of these over a previous baseline generally seem to decrease performance of the other. This is especially evident in the MED dataset, which is larger and representative

of a more diverse set of downward monotone examples (the MoNLI dataset is limited to the "No" operator). Sensibly, a decrease in performance in upward monotone contexts also leads to a decrease in performance on the original SNLI and MNLI datasets 3.6 (which are skewed in favor of upward monotone examples). However, in most cases (except for the roberta-large-mnli model on the MNLI benchmark) our method results in a *smaller* performance loss.

3.5 Conclusion and Future Work

To address **RQ 1** (*How much does fine-tuning on the HELP dataset improve NLI models' performance on existing natural logic evaluation datasets? Would a secondary transfer-learning task based on the prediction of context monotonicity result in an improvement in overall evaluation scores?*), we have presented a study of existing strategies for improvement of natural logic handling in neural NLI, together with an assembly of evaluations on previously established monotonicity reasoning evaluation sets. This is complemented with an introduction of a context monotonicity prediction task for a transfer learning improvement strategy for NLI models, treating the a *context* as an atomic unit whose monotonicity property informs the correct entailment label.

Introducing context monotonicity classification into the training pipeline of NLI models provides some performance gains on challenge datasets designed to test monotonicity reasoning. However, these are almost negligible in comparison to the performance gains from existing fine-tuning strategies. Next, we aim next to perform *structural interpretability* studies on models before and after the improvement strategies, in order to qualitatively and quantitatively compare the latent modelling of the context monotonicity feature.

More generally, we see contexts as crucial objects of study in future approaches to natural language inference. The ability to detect their logical properties (such as monotonicity) opens the door for hybrid neuro-symbolic NLI models and reasoning systems, especially in so far as dealing with out of domain insertions that may confuse out-of-the-box NLI models. The linguistic flexibility that transformer-based language models bring is too good to lose; leveraging their power in situations where only part of our sentence is in a model's distribution would be helpful for domain-specific use cases with many out-of-distribution nouns. Overall, we are interested in furthering both the *analysis* and *improvement* of emergent modelling of abstract logical features in neural natural language processing models.

3.6 Scoping and Limitations

Some other natural logic formalisms model the effect of every operator on the final entailment label (for example, the way MacCartney and Manning [23] construct large word-by-word edit sequences and have a formalism for the way word edit relation labels compose). However, we distinctly wish to move away from this level of granularity to determine (in the ensuing chapters) if the highest level of reasoning abstraction is captured, and that a "net effect" monotonicity feature is being encoded and considered. Mostly, this is because we are building towards conducting interpretability studies that can tell us more about the high-level reasoning capabilites of models. This line of enquiry requires causal diagrams and hypotheses about how features are represented. The more we decompose the reasoning chain to include all the constituent operators which give rise to the monotonicity of a context, the more complex our causal diagram would become, and it becomes harder to envision how this information may be encoded and, even more so, how to perform structural interventions for interventional studies. More generally, we see this thesis as a first step in a top-down approach to the broader investigation of whether and how NLI models can encode monotonicity. If our work finds that the high-level reasoning mechanism of context monotonicity and its interplay with concept relations is captured, then it could justify follow-up work which studies how further decompositions of the monotonicity property may be observed in NLI models.

Chapter 4

Decomposing Natural Logic Inferences in Neural NLI

This chapter presents the NLI-*XY* dataset in response to **RQ 3** (*Can we formulate and construct a natural logic dataset that is suitable for both targeted evaluation and interpretability?*), and adresses research questions **RQ 4** (*Are the intermediate features of context monotonicity and concept inclusion relations emergent in the internal representations of NLI models which perform better at natural logic tasks? Can we provide comparative quantitative evidence?*) and **RQ 5** (*Can we identify which features are responsible for errors in poorer- performing models?*) by carrying out a systematic probing study which investigates whether these models capture the crucial semantic features central to natural logic: context monotonicity and concept inclusion relations. Correctly identifying valid inferences in *downward-monotone contexts* is a known stumbling block for NLI performance, subsuming linguistic phenomena such as negation scope and generalised quantifiers. To understand this difficulty, we emphasize monotonicity as a property of a *context* and examine the extent to which models capture relevant monotonicity information in the vector representations which are intermediate to their decision making process. Drawing on the recent advances in probing practices, we compare the presence of monotonicity features across various models. We find that monotonicity information is notably weak in the representations of popular NLI models which achieve high scores on benchmarks, and observe that previous improvements to these models based on fine-tuning strategies have introduced stronger monotonicity features together with their improved performance on challenge sets.

4.1 Introduction

Large, black box neural models which achieve high scores on benchmark datasets designed for testing *natural language understanding* are the subject of much scrutiny and investigation. It is often investigated whether models are able to capture specific semantic phenomena which mimic human reasoning and/or logical formalism, as there is evidence that they sometimes exploit simple heuristics and dataset artifacts instead [7], [97].

As we have discussed in the chapter 3, behavioural studies based on targeted evaluation sets (such as in Yanaka, Mineshima, Bekki, *et al.* [13] and Richardson, Hu, Moss, *et al.* [12]) have shown that downward monotone contexts (featuring downward monotone operators such as negation markers and generalised quantifiers) result in the kinds of natural logic inferences which are often known to stump neural NLI models that demonstrate high performance on large benchmark sets such as MNLI [18].

In this chapter, we present a *structural* study: instead of looking at performance scores of the final model predictions on NLI examples (what we called *behavioural* studies in chapter 2) we wish to identify the *structural* differences between the representations of models that have poor performance accuracy on natural logic examples and those that have higher performance scores. We investigate the extent to which the features relevant for identifying natural logic inferences, especially context monotonicty itself, are encoded in the model's internal representations. To this end, we carry out a systematic *probing* study.

Our contributions may be summarized as follows:

1. We perform a structural investigation as to whether the behaviour of *natural logic* formalisms are mimicked within popular transformer-based NLI models.

2. For this purpose, we present a joint NLI and semantic probing dataset format (and dataset) which we call NLI-*XY*: it is a unique probing dataset in that the probed features relate to the NLI task output in a systematic way (following the schema in table 1.4).

3. We employ thorough probing techniques to determine whether the abstract semantic features of *context monotonicity* and *concept inclusion relations* are encoded in the models' internal representations.

4. We observe that some well-known NLI models demonstrate a systematic failure

to model context monotonicity, a behaviour we observe to correspond to poor performance on natural logic reasoning in downward-monotone contexts. However, we show that the existing HELP dataset improves this behaviour.

5. We support the observations in the probing study with several *qualitative analyses*, including decomposed error-breakdowns on the NLI-*XY* dataset, representation visualisations, and evaluations on existing challenge sets.

4.2 Related Work

Natural logic dates back to the formalisms of Sanchez [21], but has been received more recent treatments and reformulations inmaccartney-manning andccg2mono. Symbolic and hybrid neuro-symbolic implementations of the natural logic paradigm have been explored in Kalouli, Crouch, and Paiva [92], Chen, Gao, and Moss [98], and Abzianidze [99] andmonalog.

The shortcomings of natural logic handling in various neural NLI models have been shown with several *behavioural* studies, where NLI challenge sets exhibiting examples of downward monotone reasoning are used to evaluate performance of models with respect to these reasoning patterns [13], [14], [84], [100], [101].

In an attempt to better identify linguistic features that neural models manage or fail to capture, researchers have employed *probing* strategies: namely, the *diagnostic classification* [102] of auxiliary feature labels from internal model representations. Most probing studies in natural language processing focus on the *syntactic* features captured in transformer-based language models [11], but calls have been made for more sophisticated probing tasks which rely more on contextual information [74].

In the realm of semantics, probing studies have focused more on *lexical* semantics [55]: word pair relations are central to monotonicity reasoning, and thus form part of our probing study as well, but the novelty of our work is the task of classifying context monotonicity from intermediate contextual embeddings.

4.3 NLI-*XY* Dataset

In section 1.2.1, we have described a subset of NLI examples where the input premise–hypothesis pair has the form $p = f(x), h = f(y)$, and the NLI gold label relies only on the *context monotonicity* of the shared context f and the concept inclusion relation of

the noun phrase pair (x, y). We follow this structure as the basis for the NLI-*XY* dataset. This is the first probing dataset in NLP where the auxiliary labels for intermediate semantic features influence the final task label in a rigid and deterministic (yet simple) way, with these features being themselves linguistically complex. As such, it is a "decomposed" natural logic dataset, where the positive entailment labels are further enriched with labels for the monotonicity and relational properties which gave rise to them.

This allows for informative qualitative and structural analyses into natural logic handling strategies in neural NLI models: given a set of premise–hypothes input pairs (p, h), we can not only compare the model's NLI prediction output to a gold label, but we can investigate whether the intermediate feature labels (context monotonicity and concept inclusion relation) are encoded in the model's representations.

The NLI-*XY* dataset is comprised of the following (as exemplified in table 4.1):

1. A set of *contexts* f with a blank position (indicated with a lowercase 'x' or an underscore), annotated with the context monotonicity label.

2. A set of *insertion pairs* (x, y), which are either nouns or noun phrases, annotated with the concept inclusion word-pair relation.

3. A derived set of premise and hypothesis pairs $(f(x), f(y))$ made up of permutations of (x, y) insertion pairs through contexts f, controlled for grammaticality as far as possible.

			Auxilliary Label
Context	f	I did not eat any ___ for breakfast.	$\downarrow$ (downward monotone)
Insertion Pair	(x,y)	(fruit, raspberries)	$\sqsupseteq$ (reverse concept inclusion)
			NLI Label
Premise	$f(x)$	I did not eat any fruit for breakfast.	Entailment
Hypothesis	$f(y)$	I did not eat any raspberries for breakfast.	

Table 4.1: A typical NLI-*XY* example with labels for context monotonicity, lexical relation and the final entailment label.

We present examples of the component parts and their composition in table 4.1. The premise/hypothesis pairs may thus be used as input to any NLI model, while the context monotonicity and insertion relation information can be used as the targets of an auxiliary probing task on top of the model's representations.

We make the NLI-*XY* dataset and all the experimental code used in this work is publically available [1]. We constructed the NLI-*XY* dataset used here as follows:

Context Extraction We extract context examples from two NLI datasets which were designed for the behavioural analysis of NLI model performance on monotonicity reasoning. In particular, we use the manually curated evaluation set MED [13] and the automatically generated HELP training set [14]. By design, as they are collections of NLI examples exhibiting monotonicity reasoning, these datasets mostly follow our required $(f(x), f(y))$ structure, and are labelled as instances of upward or downward monotonicity reasoning (although the contexts are not explicitly identified).

We extract the common context f from these examples after manually removing a few which do not follow this structure (differing, for example, in pronoun number agreement or prepositional phrases). We choose to treat determiners and quantifiers as part of the context, as these are the kinds of closed-class linguistic operators whose monotonicity profiles we are interested in. To ensure grammatically valid insertions, we manually identify whether each context as suitable either for a singular noun, mass noun or plural noun in the blank/"*x*" position.

Insertion Pairs Our (x, y) insertion phrase pairs come from two sources: Firstly, the labelled word pairs from the MoNLI dataset [84], which features only single-word noun phrases. Secondly, we include an additional hand-curated dataset which has a small number of *phrase-pair* examples, which includes intersective modifiers (e.g. ("brown sugar", "sugar")) and prepositional phrases (e.g. ("sentence", "sentence about oranges")). Several of these examples were drawn from the MED dataset. Each word in the pair is labelled as a singular, plural or mass noun, so that they may be substituted only into those contexts which allow for a resulting sentence which is grammatically valid.

Premise/Hypothesis Pairs Premise/Hypothesis pairs are constructed by substituting the insertion pairs in all possible combinations through the set of contexts within

[1] We include our dataset and experimental code at `https://github.com/juliarozanova/nli_xy`

the grammatical constraints. Such a substitution strategy may generate examples which are not consistently *meaningful*, but we see the monotonicity reasoning pattern as sufficiently rigid and syntactic that it is of interest to observe how models treat less "meaningful" entailment examples that still hold with respect to the natural logic formalism: for example, "I did not swim in a person" entails "I did not swim in an Irishman" at a systematic level. This does raise a question of whether we do (or even should) observe certain systematic behaviours on out-of-distribution examples: we leave the further investigation of this matter for future work.

Lastly, we note that the data is split into train, dev and test partitions *before* this permutation occurs, so that there are *no shared contexts or insertion pairs* between the different data partitions, in an attempt to avoid overlap issues such as those discussed in [103]. The full dataset statistics are reported in table 4.2.

		Context Monotonicity		
Partition	**(x,y) Relation**	Up ↑	Down ↓	Total
train	$\sqsubseteq$	671	543	1214
	$\sqsupseteq$	671	543	1214
	None	244	222	466
	Total	1586	1308	**2894**
dev	$\sqsubseteq$	598	389	987
	$\sqsupseteq$	598	389	987
	None	220	242	462
	Total	1416	1020	**2436**
test	$\sqsubseteq$	1103	1066	2169
	$\sqsupseteq$	1103	1066	2169
	None	502	516	1018
	Total	2708	2648	**5356**

Table 4.2: Dataset statistics for the NLI-*XY* dataset. We employ an aggressive $30, 20, 50$ train-dev-test split for a more impactful probing result, as probing is meant to demonstrate the *ease of extraction* of features. In particular, higher test accuracy with a smaller training set is a more convincing probing result than one with a large training set and small test set.

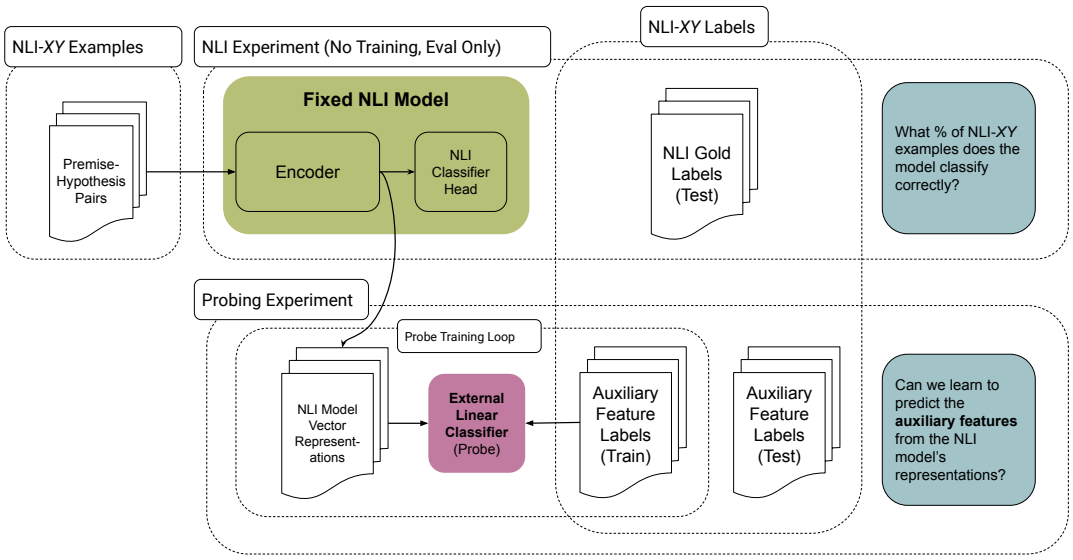


Figure 4.1: The NLI-*XY* dataset has two sets of labels: the gold labels for the NLI task and the auxiliary feature labels for the context monotonicity and the concept inclusion relation. This allows for both a standard NLI evaluation experiment and a structural interpretability experiment, allowing for external probing models to be trained to detect the intermediate features directly from a given model's representations of the input examples.

4.4 Experimental Setup

Our experiments are designed to investigate the following questions: Firstly, how do NLI models compare in their learned encoding of context monotonicity and lexical relational features? Secondly, if a model successfully captures these features, to what extent do they correspond with the model's predicted entailment label? We investigate these questions with a detailed probing study and a supporting qualitative analysis, using decomposed error break-downs and representation visualisation. We illustrate in figure 4.1 how the labels of the NLI-*XY* dataset will be used in the probing tasks (section 4.4.2) and in the standard NLI evaluation (section 4.4.3) of each model.

4.4.1 Model Choices

We consider a selection of neural NLI models based on BERT-like transformer language models (such as BERT [4], RoBERTa [5] and BART [104]) which are fine-tuned on one of two benchmark training sets: either SNLI [19] or MNLI [18]. Of particular interest, however, is the case where these models are trained on an additional dataset (the HELP dataset from [14]) which was designed for improving the overall balance

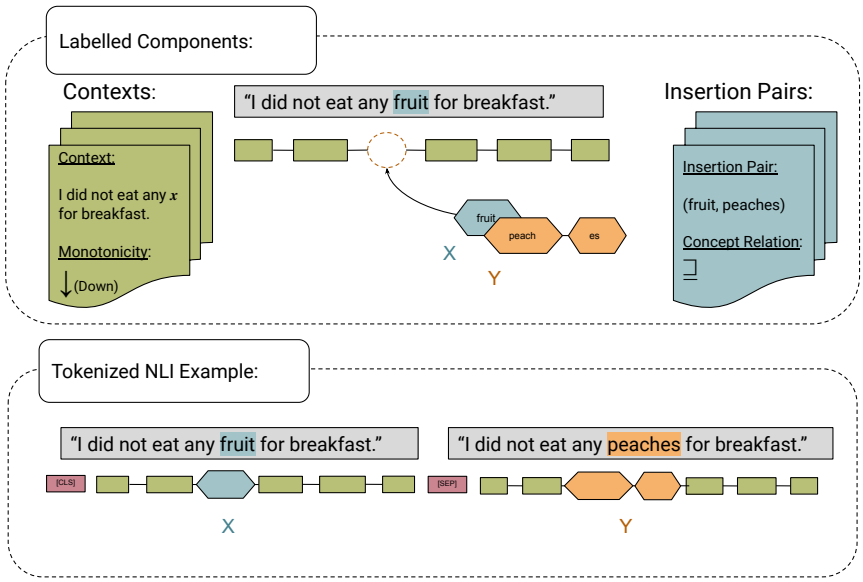


Figure 4.2: Our compositional dataset format, NLI-*XY*: the x and y token spans are the target representations we probe here.

of upward and downward monotone contexts in NLI training data. We use our own random $50-30-20$ train-dev-test split of the HELP dataset (ensuring unique contexts in every split), so that there is no overlap of contexts between the fine-tuning data and the few HELP-test examples we used as part of our NLI-*XY* dataset[2].

4.4.2 Probing Tasks

The NLI-*XY* dataset is equipped with two auxiliary feature labels which are the targets of the probing task: context monotonicity and the relation of the (x,y) word pair (referred to as concept inclusion relation or lexical relation). We now describe the details of the intermediate representations we choose as inputs to the probing tasks:

4.4.2.1 Target Representation

The standard practice for word-pair relation classification tasks is to concatenate the contextual representation vectors for the (x,y) word pair (taking the mean vector for multi-token words), as indicated in figure 4.2. We argue that this is a good representation choice for probing context monotonicity as well: as we are considering transformer-based bidirectional encoder architectures, the context (including the order) of each token

[2]We use the *transformers* library [96] and their available pretrained models for this work.

in the input sequence informs the representation of each token in the final layer. As such, we propose that since contextual information is implicitly encoded, it is feasible to expect that a token's vector representation may encode contextual features such as context monotonicity. As both the x and the y word occur in the same respective context, we are comfortable probing the concatenated (x, y) representation for contextual features, and note that it allows for easy comparison with the word pair relation probing results.

4.4.2.2 Probing Methodology

For each auxiliary classification task, we use simple linear models as probes. We train 20 probes of varying complexities using the *probe-ably* framework [105].

Probe Complexity Control The complexities are represented and controlled as follows: For a trained linear model $\hat{y} = W\mathbf{x} + \mathbf{b}$, we follow Pimentel, Saphra, Williams, *et al.* [74] in using the nuclear norm

$$||\mathbf{W}||_* = \sum_{i=1}^{min(|\mathcal{T}|,d)} \sigma_i(\mathbf{W}).$$

of the matrix W as the approximate measure of complexity. Here, $\mathcal{T}$ is the number of target classes in the probe's output, d is the dimension of the representation vectors which are inputs to the probe, and $\sigma_i(W)$ is the i-th *singular value* of the matrix W. In cases where the auxiliary task has a relatively large number of classes, the rank has been used as the proxy measure of model complexity [11]. As the nuclear norm is a convex approximation of the *rank* of the transformation matrix, it is used in Pimentel, Saphra, Williams, *et al.* [74]. This is of particular use in our case because we have a low number of prediction classes (d): two for context monotonicity and three for the concept inclusion relation. Hence, using only the rank would yield very few values, while using the nuclear norm allows for a larger number of informative values.

Accuracy and Selectivity Naively, a strong accuracy on the probing test set may be understood to indicate strong presence of the target features within the learned representations, but there has been much discussion about whether this evidence is compelling on its own. In fact, certain probing experiments have found the same accuracy scores for random representations [106], indicating that high accuracy scores are meaningless in isolation. Hewitt and Liang [64] describe this as a dichotomy

between the representation's encoding of the target features and the probe's capacity for *memorization*, and propose the use of the *selectivity* measure to always place the probe accuracy in the context of a controlled probing task with shuffled labels on the same vector representations. For each fully trained probe, we report both the test accuracy and the *selectivity* measure: tracking the selectivity ensures that we are not using a probe that is complex enough to be *overly expressive* to the point of having the capacity to overfit the randomised control training set.

Control Task The *selectivity* score is calculated with respect to a *control task*. At its core, this is just a balanced random relabelling of the auxiliary data, but [64] advocate for more targeted control tasks with respect to the features in question and a hypothesis about the model's possible capacity for *memorization*. For example, in their control task for POS tagging, they assign the same label to each instance of a word's surface form ("word type") to account for possible lexical memorization. For our context monotonicity classification control task, we assign a shared random label for all identical insertion pairs, regardless of context. Thus, a probe which is expressive enough to "memorize" context monotonicity labels associated with the examined word pairs would attain high accuracy on this control task. By construction, our context monotonicity classification task is much more context-dependent and balanced: a given (x, y) insertion will occur about as often in upward and downward monotone contexts, making it harder for a probe to exploit meaningless heuristics, such as associating a given insertion pair with a context monotonicity label. As such, we expect the selectivity scores to be low, but this is intentional and indicative of good dataset balancing. As we have no overlap of word pairs or contexts across the training and test set, we follow Pimentel, Saphra, Williams, *et al.* [74] in reporting selectivity using label shuffling on the *training* set only.

4.4.3 NLI Challenge Set Evaluations

As well as the NLI-*XY* dataset (which can function as an ordinary NLI evaluation set), for completeness we report NLI task evaluation scores on the full MED dataset [13], which was designed as a thorough stress-test of monotonicity reasoning performance. Furthermore, we report scores on the HELP-test set (from the dataset split in Rozanova, Ferreira, Thayaparan, *et al.* [107]): this data partition was not used in the fine-tuning of models on HELP, but we include the test scores here for insight.

4.4.4 Decomposed Error Analysis

The compositional structure and auxiliary labels in the NLI-*XY* dataset allow for qualitative analysis which may enrich the observations. To this end, we construct decomposed error analysis heatmaps which indicate whether a given premise-hypothesis data point $(f(x), f(y))$ is correctly classified by an entailment model. These are structured with individual (x, y) insertion pairs on the vertical axis and contexts on the horizontal axis. For brevity (and because this is representative of our observations), we include only the error breakdowns for the two sublasses of the positive entailment label: where the context monotonicity is upward and lexical relation is forward inclusion, and where the context monotonicity is downward and the lexical relation is reverse inclusion.

		Feature Probing		**NLI Monotonicity Challenge Sets**		
NLI Models	Fine-Tuning Data	Context Monotonicity (%*)	(x, y) Insertion Relation (%*)	HELP-Test (%)	MED (%)	NLI-*XY* (%)
roberta-large-mnli	-	59.00	84.00	36.69	46.10	59.01
roberta-large-mnli	HELP	84.00	76.00	**97.63**	**78.22**	**80.68**
facebook/bart-large-mnli		70.00	64.00	43.61	46.54	60.59
facebook/bart-large-mnli	HELP	73.00	70.00	88.99	77.16	79.34
bert-base-uncased-snli		73.00	51.00	63.55	49.38	49.09
bert-base-uncased-snli	HELP	73.00	51.00	66.80	46.13	44.79

Table 4.3: Summary NLI challenge test set and probing results for all considered models. *Probing results are summarized with the *accuracy at max selectivity*.

4.5 Results and Discussion

4.5.1 Probing Results

The results for the linear probing experiments for both the *context monotonicity classification* task and the *lexical relation* classification task may be found in figure 4.3, with a summary score of accuracy at maximum selectivity visible in table 4.3. The results of the control tasks are taken into account as part of the selectivity measure, which is represented on the right hand plot for each experiment.

Models that have been finetuned on HELP demonstrate the strongest context monotonicity probing scores, up to 84% for the roberta-large-mnli-help model. It is particularly notable that large models trained only on the MNLI dataset have inferior performance on context monotonicity classification. This corresponds with the further qualitative observations, suggesting that even in some of the most successful

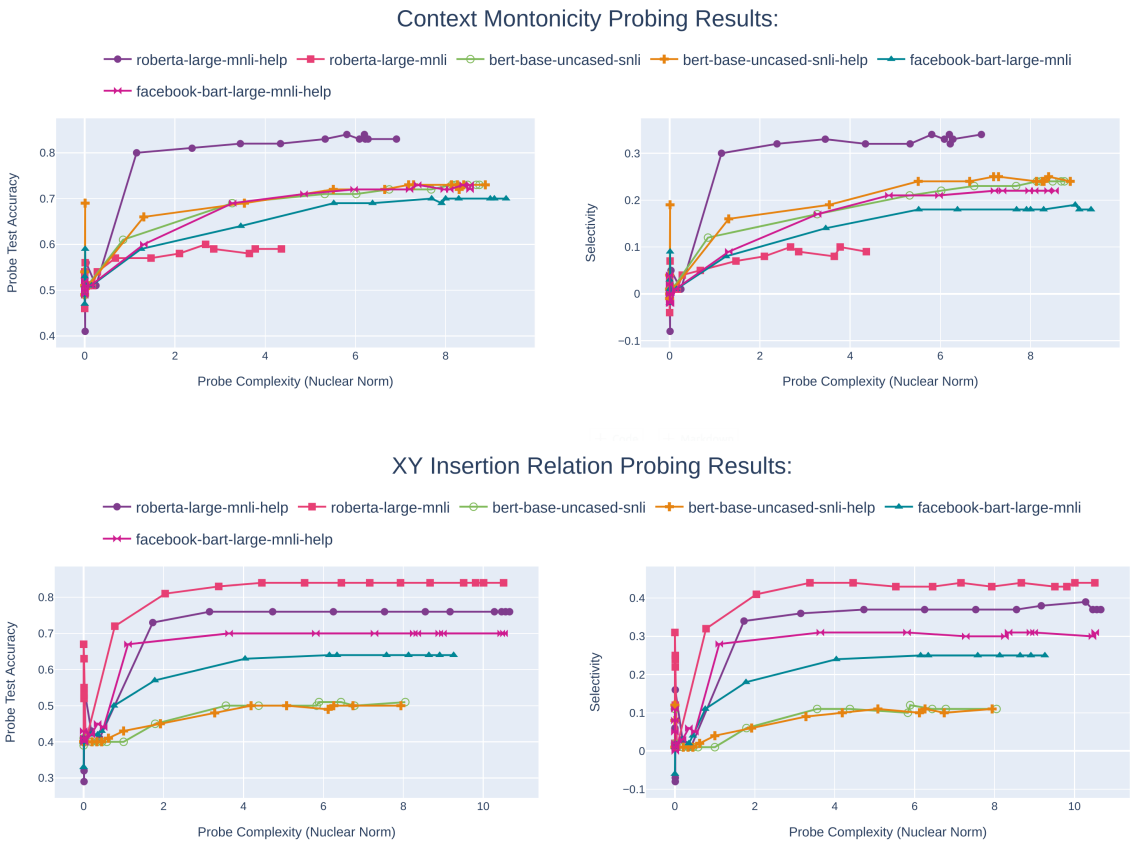


Figure 4.3: Linear probing results for all examined models.

transformer-based NLI models, *there is a poor "understanding" of the logical regularities of contexts and how these are altered with downward monotone operators.*

On the other hand, roberta-large-mnli demonstrates much higher probing scores for insertion relation classification (at 84%, higher than the second best at 76% (roberta-large-mnli-help).) Contrastingly, the version of the facebook-bart-large-mnli model finetuned on the HELP dataset shows better probing scores (peaking at around 70%) on the insertion relation prediction task than its counterpart trained only on MNLI (reaching at most 64%).

For the probing curves, we interpret probing curves which reach higher scores at lower probe complexities as better probing results in the sense that the task is in a sense "easier" to learn from the given representations.

4.5.2 Comparison to Challenge Set Performance

A summary of the probing results (presented as accuracy at maximum selectivity) can be compared with challenge set performance in table 4.3. Evaluation on the challenge test sets is relatively consistent with monotonicity probing performance, in the sense that there is a correspondence between poor/successful modeling of monotonicity features and poor/successful performance on a targeted natural logic test set. As these challenge sets are focused on testing monotonicity reasoning, this is a result which strongly bolsters the suggestion that explicit representation of the context monotonicity feature is crucial, especially for examples involving negation and other downward monotone operators. Furthermore, we generally confirm previous results that additional fine-tuning on the HELP data set has been helpful for these specialized test sets, and add to this that it similarly improves the explicit extractability of relevant context montonicity features from the latent vector representations.

4.5.3 Qualitative Analyses

Error Break-Downs An error heat map according to decomposed context monotonicity and word-pair insertion relation can be seen in figures 4.4 and 4.5. We are less concerned with the accuracy score (on NLI challenge sets) of a given model as with the behavioural *systematicity* visible in the errors, as we are not interested in noisy errors which may be due to words or phrases from outside the training domain. Consistent mis-classification for all examples derived from a fixed context or insertion pair are actually *also* strongly suggestive of a regularity in reasoning. The decomposed error analyses paint a striking picture: we generally see that models trained on MNLI routinely fail to distinguish between the expected behaviour of upward and downward monotone contexts, despite generally achieving high accuracies on large benchmark sets.

This is in accordance with observations in Yanaka, Mineshima, Bekki, *et al.* [14] and Yanaka, Mineshima, Bekki, *et al.* [13], where low accuracy on the downward-monotone reasoning sections of challenge sets points to this possibility. Howver, they show consistently show strong behavioural regularity with respect to concept inclusion. Even when the contexts are downward monotone, they still treat them systematically as if they were *upward* monotone, echoing the concept insertion pair relation *only*: they completely fail to discriminate between upward/downward monotone contexts and their opposite behaviours.

Visualisation In figures 4.7 and 4.6, each data point corresponds to an embedded example (contextual (x, y) word pair representation) in the NLI-*XY* dataset, with the left and right columns colored with the *gold* auxiliary labels for context monotonicity and concept inclusion relations respectively. These illustrate the probing observations: in the well-known roberta-large-mnli model, concept inclusion relation features are distinguishable, whereas context monotonicity is very randomly scattered, with no emergent clustering. However, the roberta-large-mnli-help model shows an improvement in this behaviour, demonstrating a stronger context monotonicity distinction.

4.6 Conclusion

In the first part of chapter 4, we address **RQ 3** (*How can we construct a natural logic dataset that is suitable for both targeted evaluation and interpretability?*) by introducing the *NLI-XY* dataset: a compositional NLI dataset with labels for the intermediate features of context monotonicity and concept inclusion relations, which make it suitable for structural interpretability studies. Furthermore, it supports qualitative error analyses which allows for checking the *consistency* with which models treat certain components while varying others. It is based on the formalism already presented in chapter 3.

In the second part of this chapter, we address **RQ 4** (*Are the intermediate features of context monotonicity and concept inclusion relations emergent in the internal representations of NLI models which perform better at natural logic tasks? Can we provide comparative quantitative evidence?*) by presenting an extensive *probing* study to determine the representation of the two semantic features relevant to natural logic, context monotonicity and concept inclusion relations. Specifically, we draw comparisons between NLI models before and after improvement strategies for natural logic handling, We have shown that state-of-the-art models seem to fail to capture the monotonicity feature, while models fine-tuned on the HELP dataset demonstrate strong emergence of this feature after some more balanced training. We confirm previous findings that state-of-the-art models do, however, show strong probing performance for the concept inclusion feature (which generalises previously studied lexical relations).

Lastly, we also contribute to the response to **RQ 5** (*Which features are responsible for errors in poorer-performing models?*) by presenting qualitative analyses in the form of *visualised projections* and informative *error breakdowns* which further bolster the argument that it is poor *context monotonicity* modelling that is a bottleneck for strong out-of-the-box NLI models' capacity for correctly identifying valid natural logic

deductions. The consistent treatment afforded to concept pairs (but not as much for contexts) by state-of-the-art models such as roberta-large-mnli complement the probing findings: the observed error heat maps are especially suggestive that monotonicity is the true bottleneck for strong natural logic reasoning here, even insofar as insertion pairs in downward monotone contexts are routinely treated as if they occur in upward monotone ones. On the other hand, the systematicity demonstrated by roberta-large-mnli-help is extremely promising: we observe that errors which arise from the mistreatment of a context or concept pair insertion are at least *applied consistently* and treated "correctly" in composition with the other reasoning component, at least with respect to its incorrect assessment. This gives hard evidence of a more principled reasoning strategy for HELP-improved models. The visualisations support both the error mapping and probing narratives, showing stronger clustering behaviour in both the key semantic features for the roberta-large-mnli-help.

In summary, the NLI-*XY* dataset has enabled us to present evidence that explicit context monotonicity feature clustering in neural model representations seems to correspond to better performance on natural logic challenge sets which test downward-monotone reasoning. In particular, the examined popular models trained on MNLI seem to lack this behaviour, accounting for previous observations that they systematically fail in downward-monotone contexts.

Furthermore, the probes' labels also have some explanatory value: both entailment and non-entailment labels can each further be broken down into sub-regions. This qualifies the classification with the observations that the data point occurs in a cluster of examples with a) upward (respectively, downward) contexts and b) a forward (respectively, backward) inclusion relation between the substituted noun phrases. In this sense, the analyses in this work can thus be interpreted as an explainable "decomposition" of the treatment of natural logic examples in neural models.

4.7 Scoping and Limitations

As our probing studies here mainly utilize *linear probes*, we can only make conclusions about linearly separable information. In particular, we cannot claim that low probing scores indicate information "absence", but merely that it is not linearly extractable. However, as the classification head is itself linearly structured, any information unaccounted for by probes cannot meaningfully be used by the task classifier anyway. Given that our end goal is to hypothesise about model reasoning strategies, we do not consider

this a critical limitation. We are more concerned with keeping our probes *simple* with as few degrees of freedom as possible, in order to strengthen as much as possible the claims that our indicated features are emergent and easily extractable.

The selectivity metric required a control task that is actually providing some redundancy in our study, as we have controlled for the same effects in our dataset construction: namely, ensuring there is no overlap between individual contexts or insertion pairs across the training and test set, so that the contextual labels may not be "memorised". However, we see it as a useful methodology that forces the consideration of confounding factors that an overly expressive probe could potentially exploit from the probing task data.

A crucial consideration that we do not address too deeply is defining our demands for *generalisation*. As mentioned briefly in chapter 4, some permutations in our compositional dataset may be less *meaningful* than others, with highly unexpected concept pairs instantiations in a given context. These are highly likely to be considered "out of domain" examples in the sense that language models may have seen no examples of, say, a certain noun being the argument of a given verb. Nevertheless, for a reasoning strategy based on a logical regularity such as monotonicity, we would expect an idealised NLI system to correctly handle unexpected instantiations, breaking the reasoning decision down to the high-level concepts of relation and monotonicity. It would be informative to, in future, compare classification success to some OOD metrics.

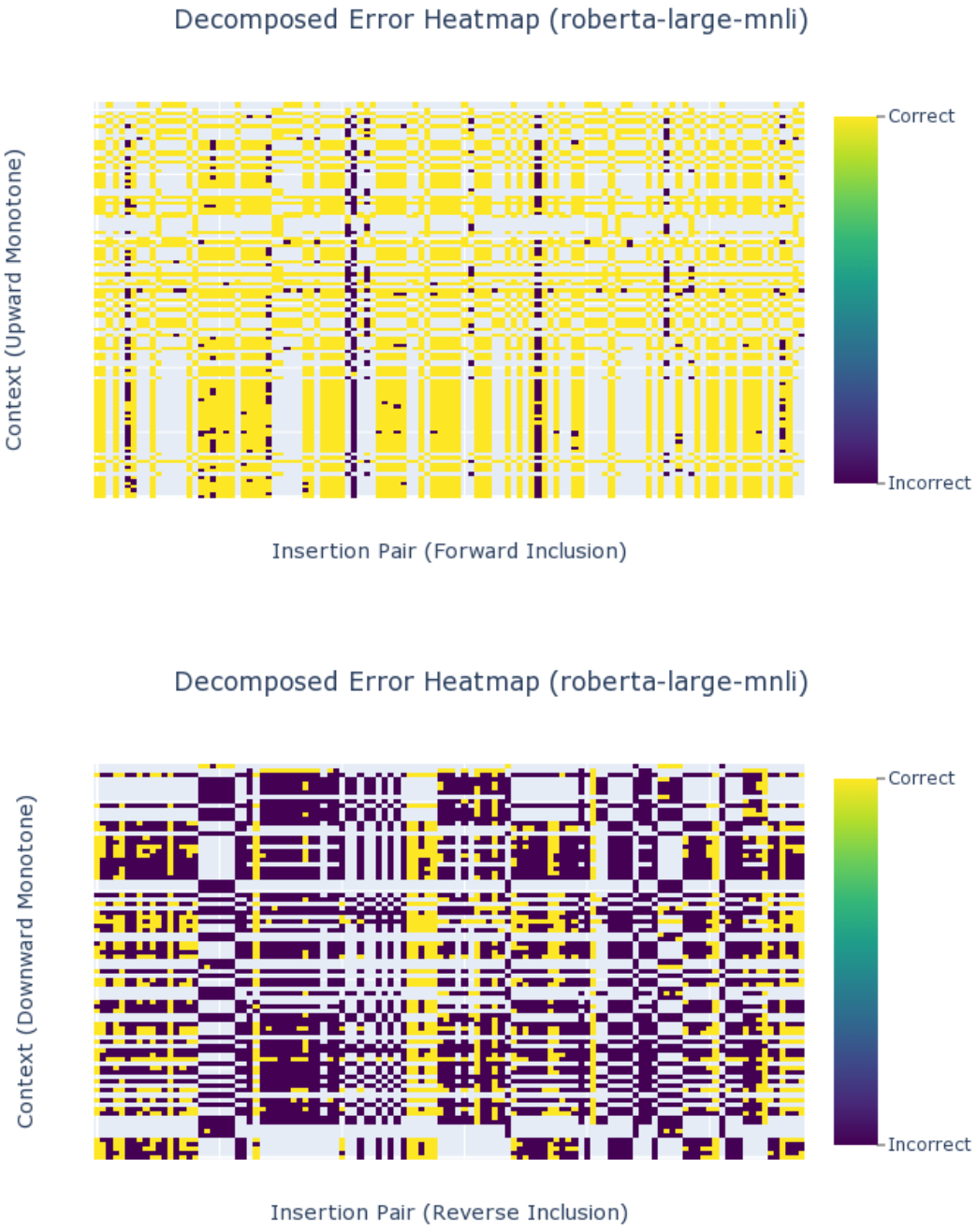


Figure 4.4: Decomposed error heat maps for roberta-large-mnli, for portions of the NLI-*XY* dataset corresponding to the indicated context monotonicity and insertion relations, expecting a postive entailment label. Individual contexts are populated along the y axis, and substituted concept pairs are populated along the x axis. Dark grid units indicate a model classification error (blank positions are present as only grammatically valid insertions were included in the dataset).

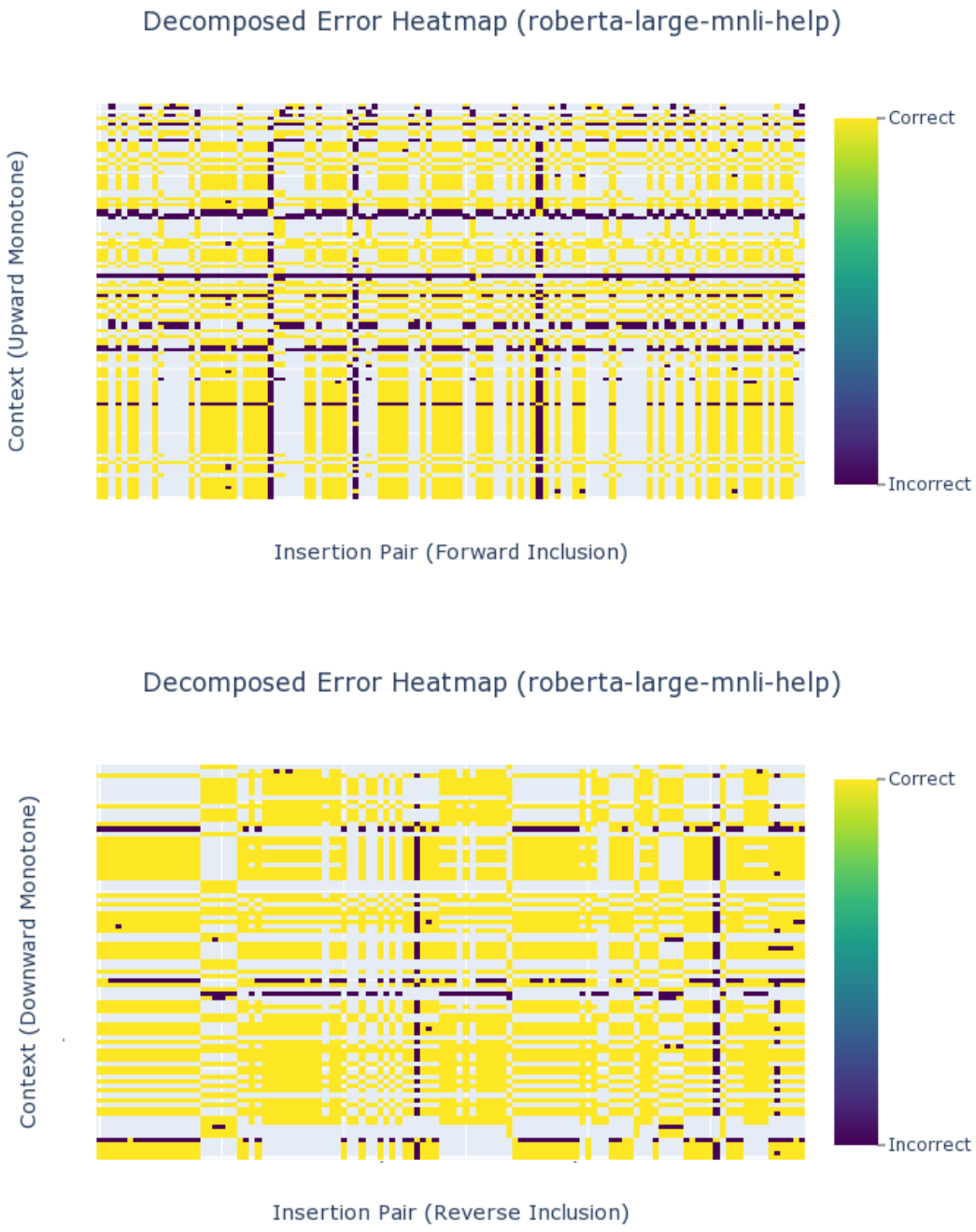


Figure 4.5: Decomposed error heat maps for roberta-large-mnli-help, for portions of the NLI-*XY* dataset corresponding to the indicated context monotonicity and insertion relations, expecting a postive entailment label. Individual contexts are populated along the y axis, and substituted concept pairs are populated along the x axis. Dark grid units indicate a model classification error (blank positions are present as only grammatically valid insertions were included in the dataset).

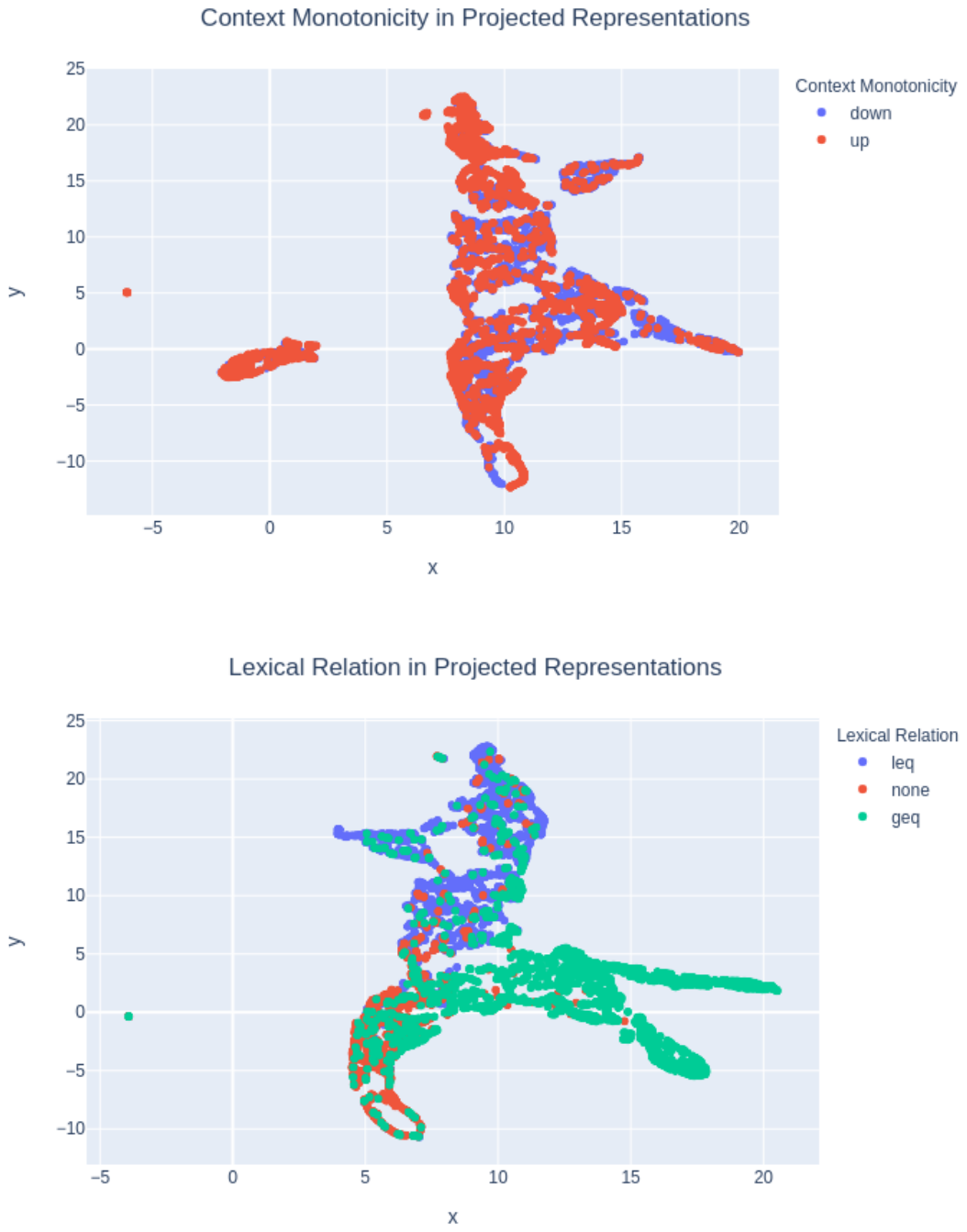


Figure 4.6: UMAP projections of concatenated (x, y) token pair representations for the roberta-large-mnli model. There is visible clustring of the concept relation/lexical relation features (in particular, a distinction between forward and reverse inclusion). However, the upward and downward monotone contexts are highly entangled.

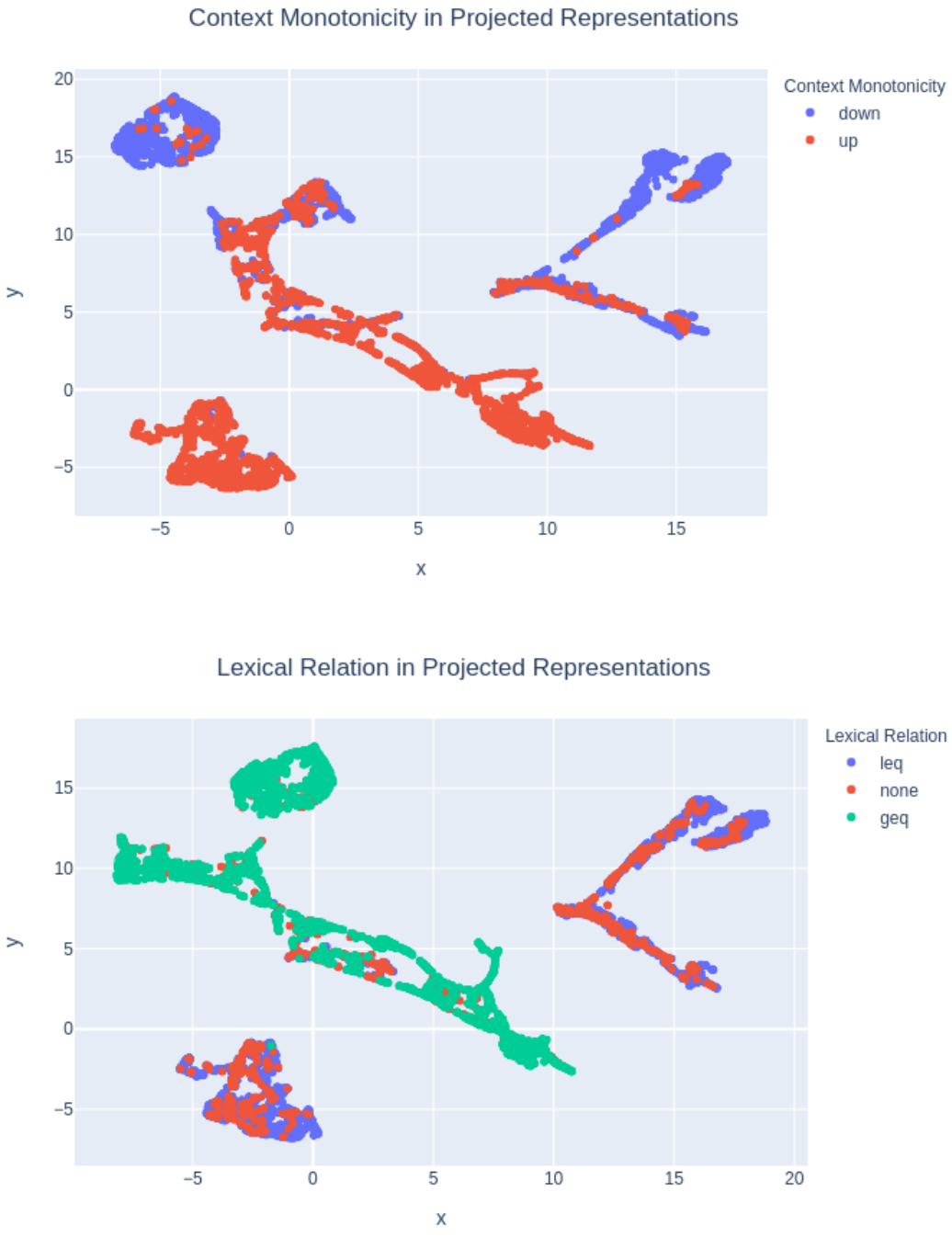


Figure 4.7: UMAP projections of concatenated (x, y) token pair representations for the roberta-large-mnli-help model, which shows greater distinction between context monotonicity features.

Chapter 5

Interventional Probing in High Dimensions

In this chapter, we adress research questions **RQ 6** (*What can structural interventional methods tell us about the usefulness of the identified representations for the NLI task?*) and **RQ 7** (*How can we devise an alternative interventional interpretability method that is still informative in the high-dimensional situations where amnesic probing fails?*) by carrying out new and existing vector-level interventions to investigate the effect of our semantic features of interest on NLI classification. We perform *amnesic* probing (which removes features as directed by learned probes) and introduce the *mnestic* probing variation (which forgets all dimensions *except* the probe-selected ones). Furthermore, we delve into the limitations of these methods and outline pitfalls that have been obscuring the effectivity of such studies.

5.1 Introduction

The *probing* paradigm has emerged as a useful interpretability methodology which has been shown to have reasonable information-theoretic underpinnings [69], [70], [73], indicating whether a given feature is captured in the intermediate vector representations of neural models. It has been noted many times that this does not generally imply that the models are *using* these learnt features, and they may represent vestigial information from earlier training steps [1], [33].

Only through interventional analyses can we start to make claims about which modelled features are used for a given downstream task: this is the aim of works such as Elazar, Ravfogel, Jacovi, *et al.* [1] and Giulianelli, Harding, Mohnert, *et al.* [61]

and Geiger, Lu, Icard, *et al.* [81]. We refer to the case where the interventions are guided by trained probes as *interventional probing*.

It has been suggested in Elazar, Ravfogel, Jacovi, *et al.* [1] (as the guidance for their *amnesic probing* methodology) that if features are strongly detected by probes, one may use debiasing methods such as *iterative nullspace projection (INLP)* [78] to intervene on the corresponding vector representations and effectively "remove" the features before re-insertion into the given classifier. Investigating the effect of these intervention operations on the classifier performance could allow for stronger causal claims about the role of the probe-detected features.

In this work, we delve deeper into the amnesic probing methodology with an NLI case study and identify two key limitations. Firstly, there is an issue of dimensionality: when the number of dimensions is high and the number of auxiliary feature classes is low, it seems that amnesic probing is not sufficiently informative. In particular, we cannot rely on the same control baselines to reach the kind of conclusions discussed in [1], as nulling out small numbers of random directions consistently has no impact on the downstream performance. Secondly, in the linguistic settings explored in Elazar, Ravfogel, Jacovi, *et al.* [1], we do not have expectations for exactly *how* or even *if* the explored features should be affecting the downstream task. This makes it difficult to explore the effectivity of the methodology itself.

To this end, use the *natural logic* subset of NLI defined in section 1.2.1. In this setting, the intermediate linguistic feature labels for *context montonicity* and *lexical relations* are already known to be extractable to a high degree of accuracy from certain NLI models' hidden layers with linear probes [108], allowing us a certain amount of understanding and control of these features' representations in the latent space. Using the deterministic and well-understood nature of the problem space where we have concrete *expectations* about the theoretical interaction between the intermediate features and the downstream label, we may critically analyse the effectivity of interventional probing.

Through the application of probe-based interventions in this setting, we show that blindly applying the amnesic probing argument structure leads to unexpected and contradictory conclusions: the two features which the final label is *known* to depend on are shown to have no influence on the final classification (both jointly and independently). This further calls into question the suitability of these methods for situations where a small number of feature label classes and high dimensionality of representations is concerned. Even more perplexingly, when we treat the NLI gold

label itself as an intermediate feature which can be nulled out with INLP, we yet again observe *almost no change to the NLI performance*. As such, the feature removal strategy appears ineffective here: we attribute this to the disproportionate size of probe-selected feature subspaces to the very high-dimensional representations.

In response, we introduce and study a variation which we call *mnestic* probing, which we show to be more informative in the high-dimensional, low-class-count setting: the core idea is to *keep only* the directions identified by the iteratively trained probes. This allows us to analyse much lower dimension subspaces, while making better use of the outputs of the INLP strategy used in amnesic probing.

We find that *mnestic probing* leads to more informative observations which are a) in line with expected behaviour for natural logic, and b) yield results which seem to better discriminate between model behaviours.

In summary, the contributions of the paper are as follows:

1. We propose the setting of *natural logic* to be ripe territory for exploration of interventional probing strategies.

2. We note two limitations of the amnesic probing methodology, demonstrating both dimensionality limitations for the control baselines 5.4.4 and contradictory behaviour in the NLI setting 5.4.2 (namely that that the expected effects of semantic features on the downstream NLI task are notably absent).

3. Building upon previous interventional methodologies, we introduce an additional *mnestic* intervention operation which uses the outputs of the INLP process in the opposite way.

4. We contrast the mnestic probing strategy with the amnesic probing results, and demonstrate it presents more informative results which are aligned with the constructed expectations in our high dimensional, low label class count setting.

5.2 Interventional Probing

We may summarise the general setup of interventional probing as follows: suppose we start with a classification model that may be decomposed as $f \circ g : \mathcal{X} \to \mathbb{R}^n$, where g is an encoder module which yields a representation which serves as an input to the classifier head f, and n is the number of output classes of the final classifier. We aim to

intervene on the output of g and observe the change in the performance of f (usually in comparison with some kind of random control baseline intervention).

Linear probes (also known as *diagnostic classifiers*) are able to identify subspaces in which a given intermediate feature set is found to be represented. These may be used as a guide for vector-level interventions on the representation space; we are specifically concerned with interventions which are vector *projections*. Otherwise, the exact nature of this intervention is interchangeable. We consider two projection strategies in particular: the *amnesic* intervention introduced in Elazar, Ravfogel, Jacovi, *et al.* [1] (described further in section 5.2.2) and our *mnestic* variation which uses the same INLP technique (section 5.2.3).

5.2.1 What Should it Tell Us?

The interventional probing steps are performed on exactly the representation that would have been an input to the classifier head f. We may re-insert the intervened representations and re-calculate the classifier accuracy (note that the iterative projections in sections 5.2.2 and 5.2.3 maintain the original dimensionality of the vector set but reduce the *rank*).

We are looking to see if the downstream performance of the classifier f drops. If it does, the interventions have removed information that was necessary for successful classification. However, as any projection would remove some information, these results must be viewed in the context of a control intervention: if the INLP process ends up removing n directions, a sample of n randomly chosen directions is selected from the original representation. Elazar, Ravfogel, Jacovi, *et al.* [1] argue that if the amnesic downstream performance drops significantly more than the random removal control performance, we may conclude that the features were necessary for the final downstream classification. On the other hand, if the performance does not drop at all, the features were not useful for the classifier in the first place. In the ensuing sections and results, we demonstrate that this is not necessarily a valid conclusion.

5.2.2 The Amnesic Intervention

We follow the procedure in Elazar, Ravfogel, Jacovi, *et al.* [1] (in turn based on *iterative nullspace projection* in Ravfogel, Elazar, Gonen, *et al.* [78]): given a fixed set X of encoded representations for the textual input (with dimensions `embedding_dimension` $\times$ `num_examples`, where the latter is the number of input data points).

We start an iterative process by training an initial linear SVM classifier to predict a set Y of auxiliary feature labels from the set of data representations X with as high an accuracy as possible. Let W_0 denote the trained weight matrix of this linear classifier, where the vectors of W_0 define directions onto which the probe projects the representations for auxiliary label classification (i.e., these are the chosen directions most aligned with auxiliary class separation). W_0 has the dimensions `num_classes` $\times$ `embedding_dimension`, where `num_classes` is the number of prediction output classes for the trained linear classifier. We wish to create a projection matrix P_0 which satisfies

$$W_0(P_0X) = 0,$$

(namely, projection onto the nullspace of W_0) which would mean that the labels Y cannot be distinguished from the projected representations P_0X using the classifier given by W_0. To do this, let R_0 donate the matrix which is the projection matrix onto the rowspace of W_0, namely

$$R_0 := W_0^\top (W_0 W_0^\top)^{-1} W_0.$$

The matrix

$$P_0 := (I - R_0)$$

is the projection matrix onto the nullspace of W_0, the orthogonal complement of the rowspace of W_0, and this projection matrix gives us a new set of representations $P_0(X)$ (the same size as X) from which the classifier W_0 cannot predict the labels Y.

We continue the iterative nullspace projection process by training *new* linear classifier weights W_i by learning to predict the auxiliary features labels Y from the projected representations $P_{i-1}X$. Now, we wish to create a projection matrix P_i which satisfies

$$W_i(P_i(P_{i-1}(X))) = 0.$$

Once again, let R_i donate the matrix which is the projection matrix onto the rowspace of W_i, namely

$$R_i := W_i^\top (W_i W_i^\top)^{-1} W_i.$$

The matrix

$$P_i := (I - R_i)$$

is the projection matrix onto the nullspace of W_i, the orthogonal complement of the rowspace of W_i, and this projection matrix gives us a new set of representations $P_i(P_{i-1}(X))$

from which the classifier W_i cannot predict the labels Y.

This process is repeated until new classifiers W_i are no longer able to achieve a task accuracy higher than the random basline for the auxiliary feature prediction task. Let W_n denote the final classifier trained in this way. The ultimate aim of the INLP process is to create a projection matrix P which *simultaneously* satisfies $W_i(PX) = 0$, for every $i \in \{0...n\}$. This is done by defining the matrix P to be the projection onto the intersection of all of the nullspaces of each W_i, but in practice we follow Ravfogel, Elazar, Gonen, *et al.* [78] in using the result of Ben-Israel [109], which shows that we can equivalently define P as the orthogonal complement of the projection matrix onto the union of the rowspaces: hence, the projection matrix P to the intersection of the nullspaces is defined to be the matrix

$$P := (I - (R_0 + ... + R_n)).$$

The matrix product PX is the projection of the data X which results in a matrix in the original dimensions of X, but with its rank reduced by the number of iteration steps (as each projection "flattens out" the representation in these directions). Projection to the intersection of nullspaces is thus the removal of any information pertaining to the auxiliary feature labels (or at least, the information which allows high performance for a linear probe): the trained linear classifiers can no longer distinguish the feature labels from these modified representations. The resulting representation PX is treated as an altered representation where this feature is *removed* or forgotten.

5.2.3 A Variation: The Mnestic Intervention

Elazar, Ravfogel, Jacovi, *et al.* [1] perform a series of experiments on various linguistic features which had previously been shown to be well-captured in language model representations and use the amnesic probing methodology to distinguish between features that are *used* by the model and those that are not by comparing post-intervention downstream task performance to a baseline of randomly removed directions.

Rather than projecting the embedded representations to the intersection of nullspaces of the trained probes (removing the target property), we project them to the *union of the rowspaces* with the transformation:

$$\begin{aligned}(I-P)X &= (I-(I-(R_0+\ldots+R_n)))X \\ &= (R_0+\ldots+R_n)X\end{aligned}$$

This has the opposite effect: we use projection to null out *everything except* the directions identified by the probes as indicative of the target feature. As such, we "remember" only the part of the representation that is predictive of that feature rather than forgetting it.

5.3 Experimental Setup

In this study, we use interventional methods [1] to study the internal behaviour of NLI models. We compare amnesic and mnestic variations of the INLP strategy, evaluating intermediate feature probing performance and downstream NLI performance after every step of the intervention process. For each auxiliary feature label and and model, we perform the *interventional probing* strategy as outlined in figure 5.1.

5.3.1 Dataset

Our setting for this study the subset of NLI examples defined in section 1.2.1, specifically using the use the NLI-*XY* dataset from chapter 4.

By construction, the NLI-*XY* dataset consists of NLI examples which rely on exactly these two abstract features: context monotonicity and the concept inclusion relation (or lexical relation) of the substituted terms. We perform two flavours of probe-based interventions (described fully in section 5.2) with four feature label sets (described next).

Auxiliary Feature Labels We begin with the two relevant intermediate features (respectively, context monotonicity and lexical relation) which are already known to correlate with stronger performance on the downstream NLI-*XY* task [108]. We will refer to this as *single-feature* interventional probing, as the probing and intervention steps are only applied to one feature set at a time. Next, we combine the two features in a cross product, creating a new feature label set with all possible combinations of these intermediate features (in the dataset, they are completely independent variables by construction [110]). We refer to this as the *composite feature label*.

Lastly, we also consider the *entailment label* itself (the downstream task label) as an input to the interventional probing process. The latter is particularly useful as a diagnostic sanity check, and aids the critical nature of our findings.

[1] We reuse much of the code included with [1], but we include our data and reproducible experimental code at `https://github.com/juliarozanova/mnestic_probing`.

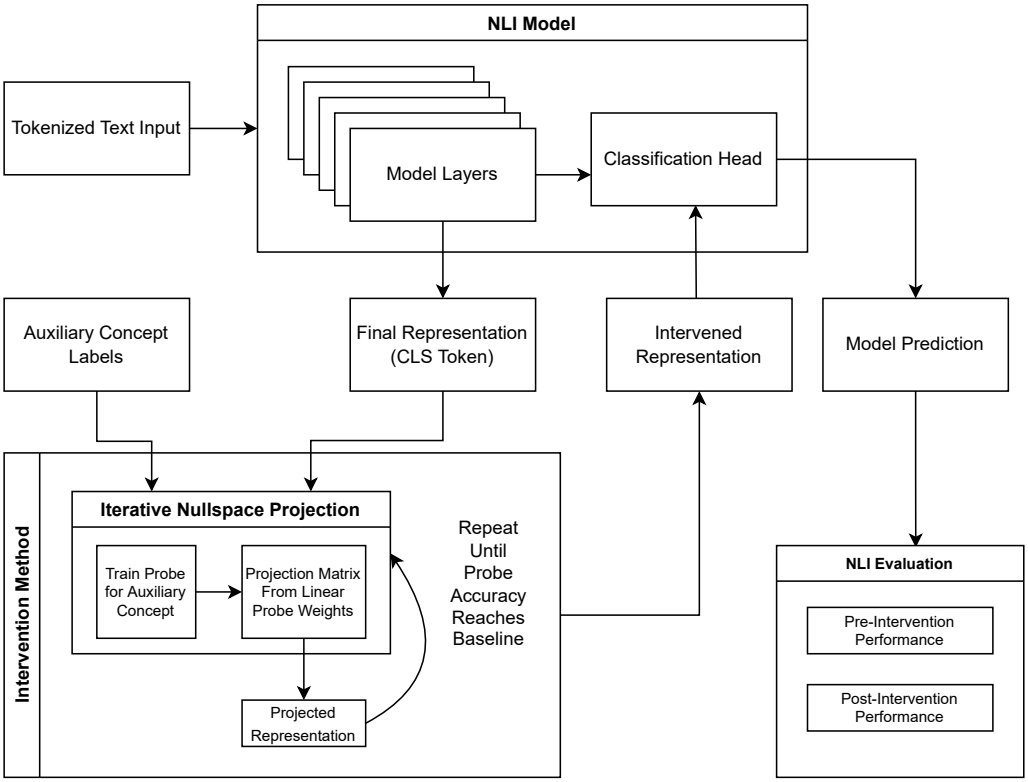


Figure 5.1: Workflow for interventional probing for NLP classification models: a basis for both the *amnesic* and *mnestic* intervention strategies.

5.3.2 NLI Models and Encoding

We compare a selection of BERT [4] and RoBERTa [5] models trained for NLI classification. Firstly, we include a pair of models trained respectively on the MNLI [18] and SNLI [19] benchmark datasets. In [108] and [110], it is shown that when roberta-large-mnli (a model which performs well on benchmarks but poorly on the targeted NLI-*XY* challenge set) receives additional training on the adversarial HELP dataset [14] it improves in NLI-*XY* performance and *begins to show high probing performance for the relevant intermediate features*, context monotonicity and lexical relations: this is the necessary precondition for doing interventional probing. We include two of their models with this property: roberta-large-mnli-help and roberta-large-mnli-double-fine-tuning (the model additionally trained on the HELP-Contexts dataset in chapter 3), with the other models included for a contextual comparison.

We perform probing and intervention on the final representation that precedes the NLI classification head: in the case of BERT and RoBERTa, this is the `[CLS]` token of the final layer.

The initial input is a tokenized NLI example from the NLI-*XY* dataset. The findings in [108] show that the intermediate feature labels (context monotonicity and lexical relations) are detectable in the concatenated tokens of the substituted noun phrases: however, for interventional purposes, we perform the probing and intervention steps on the `[CLS]` token which serves as an input to the NLI classifier head: we have found that the same features are detectable to a comparable standard, and this is the only position at which we are able to make a sensible intervention that would allow conclusions about the final classifier head only.

5.3.3 Evaluation

The significant metrics for these interventional probing paradims are the *probing accuracy* before and after the iterative nullspace projection steps (a decline to random performance indicates the feature is being "removed" from the representation in the sense that it is no longer detectable by linear probes) and the *downstream classification accuracy* on the NLI task the model's were trained for (in our case, we report the accuracy on the NLI-*XY* task).

For amnesic probing, we report the performance deltas for both the probing and downstream tasks. However, for mnestic probing, a slightly more nuanced and qualitative view is helpful: it can be assumed that eventually mnestic probing will reach

		Probing Performance		NLI-*XY* Performance	
Model	**Feature**	**Start**	**Intervention Δ**	**Start**	**Intervention Δ**
roberta-large-mnli-help	insertion relation	80.58	-40.35	79.79	0.06
	context monotonicity	87.65	-46.22	79.79	-0.09
	composite	64.48	-43.95	79.79	0.32
	entailment label	78.05	-37.49	79.79	-1.57
roberta-large-mnli-double-fine-tuning	insertion relation	62.7	-36.49	80.04	0.11
	context monotonicity	89.79	-43.28	80.19	0
	composite	57.64	-49.56	80.08	-1.67
	entailment label	82.8	-24.94	80.19	-16.53
roberta-large-mnli	insertion relation	80.39	-45.59	57.22	8.99
	context monotonicity	75.44	-27.49	57.37	-0.43
	composite	72.35	-53.51	57.24	-2.27
	entailment label	73.6	-15.31	57.37	0.1
bert-base-uncased-snli-help	insertion relation	59.53	-19.1	45.95	0.28
	context monotonicity	82.72	-33.94	45.52	-2.35
	composite	37.19	-17.08	45.76	13.68
	entailment label	47.05	0.38	45.91	0
bert-base-uncased-snli	insertion relation	60.26	-35.14	48.99	1.05
	context monotonicity	81.09	-30.77	49.42	-6.25
	composite	35.37	-17.83	50.73	7.45
	entailment label	42.44	-0.24	49.42	0

Table 5.1: Amnesic probing performance deltas across models and target feature labels: first listed is the performance on the probing task with respect to the indicated feature, and then the accuracy on the downstream NLI-*XY* task. We note the results pre-intervention and the ensuing change in accuracy.

comparable performance to the untouched vector representations, but we are interested in the comparative rates at which this happens. As the interventions are iterative, we may feed the intervened representations into the classifier head at *each step* of the intervention process —we use this to provide a step-wise presentation of results in linear plots in figure 5.4.

While the tabulated deltas in table 5.1 results are sufficient to present our observations on amnesic probing, for comparison we also include the stepwise graphical presentations in section 5.9, based on the appendix of the original publication.

5.4 Results and Discussion

5.4.1 Single Feature Amnesic Probing

The results for the standard amnesic probing procedure are in table 5.1. In particular, the single feature results are in the rows with features labelled *insertion relation* and *context monotonicity*. The amnesic operation is successful – the respective probing

accuracies approach and reach the majority class baseline.

We also include the step-wise plots of both probing performance and downstream NLI task performance: we single out the case of the insertion relation label in figures 5.2 and 5.3, but include the full suite of expanded plots for each feature in figure 5.10. The length of the iterative amnesic probing process is indicative of the number of dimensions removed to reach this baseline: it can also be considered a proxy for the strength of the feature presence in the representations, or rather, the dimension of the semantic subspace corresponding to the target features.

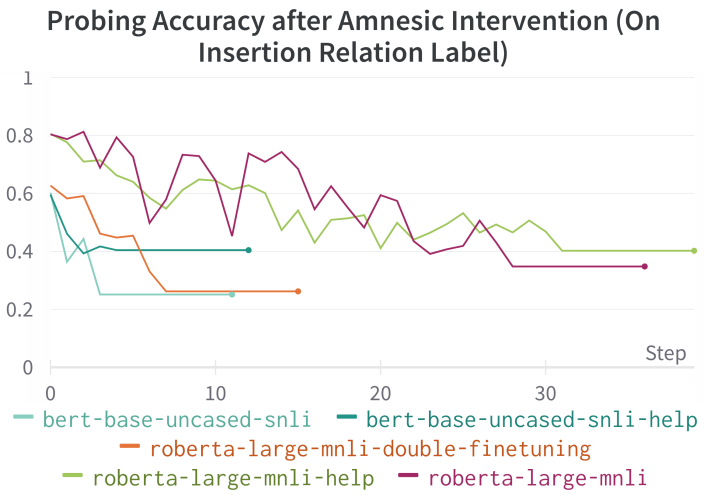


Figure 5.2: Step-wise probing performance throughout the amnesic probing process: a decrease towards the random baseline accuracy (roughly 0.3 for this 3-class task) indicates the feature is less and less extractable from the remaining representations as the iterative process continues.

The second phase of this process, i.e. the resubstitution of the modified representations as inputs to the NLI classifier head, can be seen in the right hand portion of table 5.1, labelled *NLI-XY Performance*. The result is unexpected: for each of these features, *the downstream task performance appears to be unaffected after their removal.* This is surprising when the dataset is explicitly controlled to rely only on these two features.

5.4.2 Multi Feature Amnesic Probing

The results for the amnesic probing procedure utilizing *both* auxiliary feature label sets and the entailment gold label are in the rows of table 5.1 with the labels *composite* and *entailment label* respectively. More detailed results providing accuracy scores for

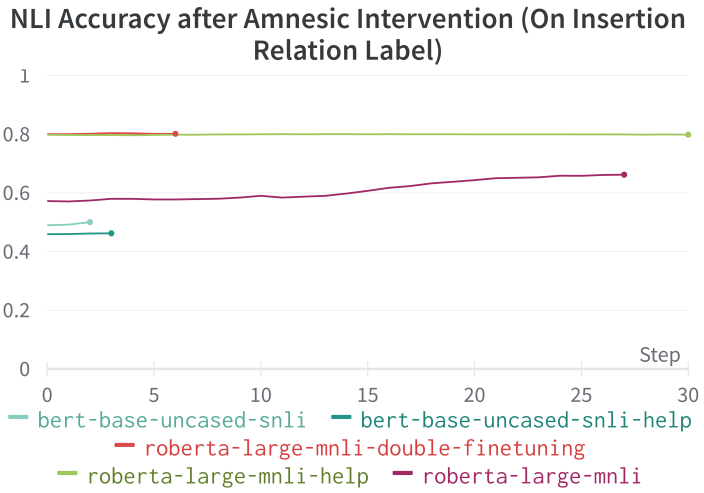


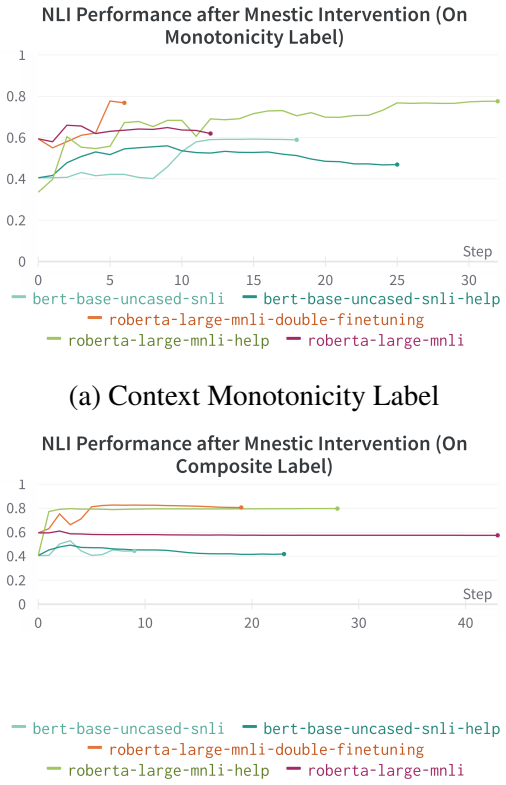
Figure 5.3: Downstream performance on NLI-*XY* after amnesic intervention (removing lexical relation information). For such an important feature to the end-task, we would expect to see a drop: but we don't!

every step of the INLP process can be seen in figures 5.11 and 5.12. We observe that once again, the downstream task performance is mostly unaffected. Unlike the unexpected result in the previous section, it is difficult to argue away the fact that this is somewhat contradictory: while single feature removal may be subject to some confounding bias, the removal of both features exhausts the variables on which this classification depends. This is highly unexpected, and suggests a point of failure for the amnesic probing process. Naturally, we cannot be without doubt that despite all our best efforts to work with a controlled dataset that relies only on these two known (but still complex) features, a model may yet find unrelated heuristics to exploit that may correlate so strongly with the downstream task label that it may perform well without representing and using these intermediate features. However, we imagine it to be a rather low probability scenario that the model learns such heuristics while simultaneously learning representations that create strongly clustered regions in the representational space for the known intermediate features *without using them at all*. The models which we have observed to perform more or less well on NLI-*XY* (such as roberta-large-mnli) are indeed estimated to be using sub-par heuristics, but this also comes with poor probing results for the intermediate features – naturally, this in itself does not imply anything conclusive, but certainly adds to our convictions.

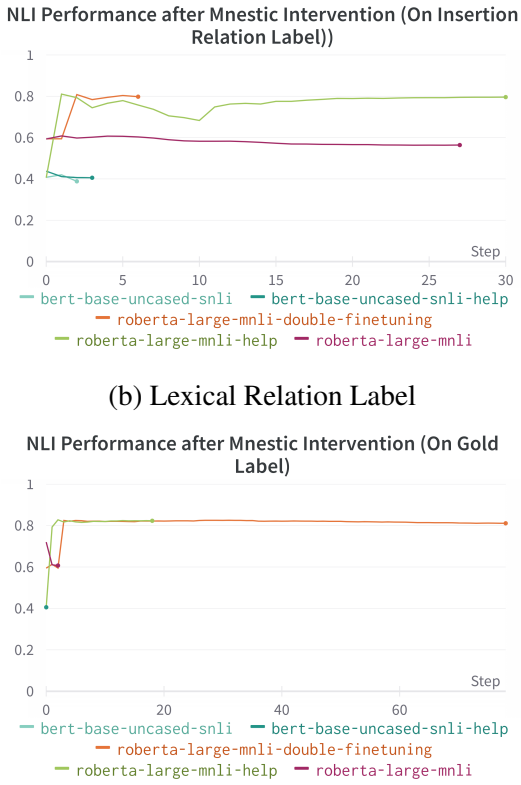
On a seprate note, it is noted in Elazar, Ravfogel, Jacovi, *et al.* [1] that there is no control for the number of dimensions removed, while there is a clear correlation

between downstream task performance and the number of label classes (and thus removed probe directions). Our feature sets have only 2 and 3 classes respectively. In the most analagous result in [1] where the auxiliary features had very few classes and no change on the downstream performance was observed, it was concluded that the features must have no effect on the outcome. It is very likely that *too little information* is being removed in this process to observe any impact on the downstream task performance. This could potentially be pointing to high redundancy in the representations which the amnesic intervention may struggle to remove appropriately.

5.4.3 Mnestic Probing



(a) Context Monotonicity Label



(b) Lexical Relation Label

(c) Composite Label

(d) Gold Label

Figure 5.4: Downstream NLI task performance after mnestic interventions, with the input representations from the corresponding step of the INLP process.

Given the possible dimensionality problem, the alternative method of *mnestic* probing seems promising: after the mnestic intervention, many dimensions are removed and few remain, so it appears to be a ripe setting for observing and comparing effects on

downstream NLI accuracy at a finer granularity. The results for NLI-*XY* task accuracy after the *mnestic* probing procedure are presented as step-wise plots in figure 5.4. There is a clear increase in NLI performance with subsequent addition of probe-chosen directions to the representations, especially viewed in the context of section 5.4.4, where we compare the performance to random choices of included directions. In the latter, performance varies randomly rather than presenting a structured increase as seen here.

We observe that the *composite* label and the gold *entailment* label are reflected in line with expectations in the mnestic probing experiments: the inclusion of the probe-selected dimensions with respect to these labels introduces a sharp and immediate increase in the NLI classifier performance. This is significantly steeper than the baseline increase observed in random addition of representation directions. Similarly, the increase is nearly as sharp for the lexical relation label. However, although an increase is observed during the iterative mnestic probing intervention for context montonicity, this increase is not at a dramatically higher rate than adding subsequently more directions from the original representation. For monotonicity specifically, this is not enough to conclude that the feature (or at least, the corresponding probe-selected dimensions) are critical to the final classifier. Nevertheless, we have been able to make clearer observations than were possible in the amnesic probing setting.

5.4.4 Control Comparison

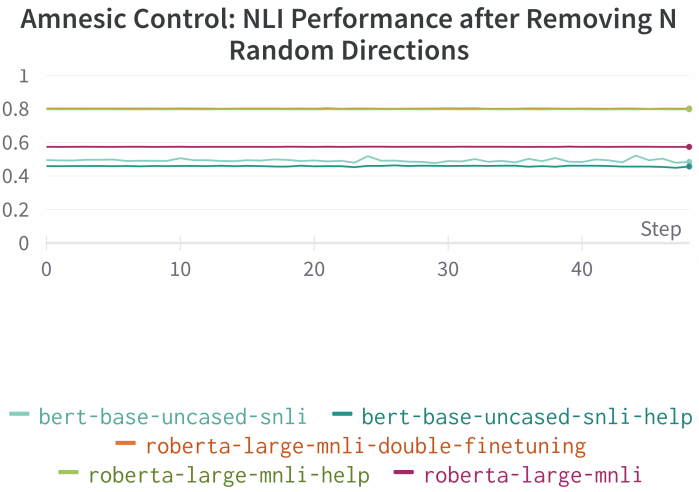


Figure 5.5: Amnesic control experiment: downstream NLI accuracy upon the *removal* of n random directions of the original representation.

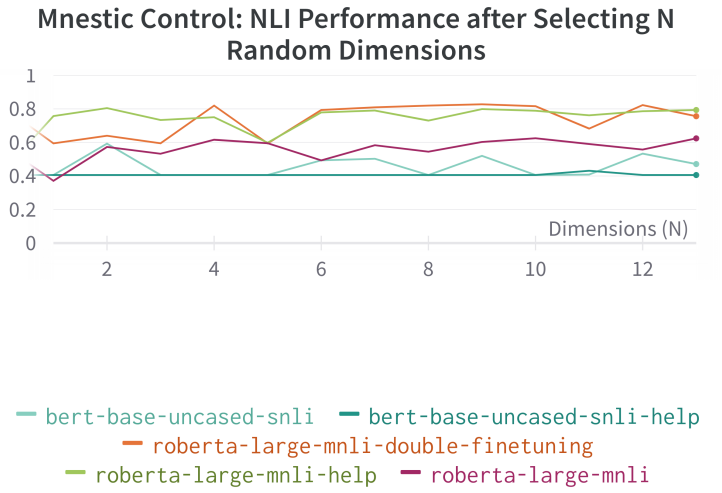


Figure 5.6: Mnestic control experiment: downstream NLI accuracy upon the *selection* of n random directions of the original representation.

We contextualise all the preceding results with a set of control experiments both for amnesic (figure 5.5) and mnestic (figure 5.6) probing. Note in particular that even with very few random dimensions kept, downstream performance starts approaching comparable levels to the full representations. As such, a single random baseline as in Elazar, Ravfogel, Jacovi, *et al.* [1] can be misleading: there is enough variability in the random direction results so as to allow for a false claim of feature irrelevance by simply getting lucky; as few as 3 dimensions can perform at the original model's performance level or arbitrarily lower.

Lastly, we compare to the mnestic probing results in figure 5.4: with the probe-selected mnestic dimension choices, the increase in downstream performance does seem to happen faster and in a more consistent fashion, while the selection of n randomly chosen directions introduces very haphazard performance spikes. This suggests the probe-selected dimensions are consistently adding to the model's access to the relevant information, and this may be stronger evidence for the usefulness of the examined features for the final classification.

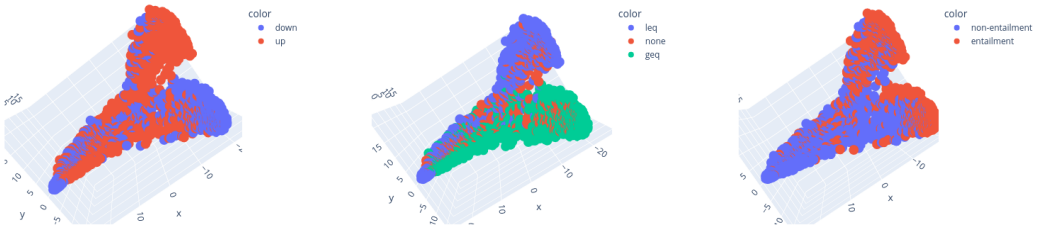
5.5 Qualitative Visualisations

For our best-berforming NLI-*XY* model (roberta-large-mnli), we include some visualisations of projected dataset representations which provide some interesting insights into the structure of representations before and after the interventional steps. We project the

computed representations of the [CLS] tokens for the NLI-*XY* test set using the UMAP algorithm of McInnes, Healy, and Melville [111]. We project onto three dimensions and present certain angles here which we found informative. Naturally, the nature of the observations here are purely speculative and do not necessarily constitute hard evidence of model behaviour, but they have certainly complemented the experimental findings.

5.5.1 Visualisations of Unmodified Representations

We first present visualisations of the unchanged representations, with each data point coloured respectively according to the context monotonicity label, the concept inclusion relation label and the final entailment label. All the reasoning components are visibly clustered, and it is interesting to note that the two prominently emergent subregions of the entailment region in figure 5.7 (c) correspond to regions which are jointly upward monotone and exhibit the forward concept inclusion relation (labeled "leq"), or are both downward monotone and in the region for the reverse concept inclusion relation (labeled "geq").



(a) Context Monotonicity Label (b) Insertion Relation Label (c) Entailment Label

Figure 5.7: Projected representations of the [CLS] tokens for the NLI-*XY* test set (specifically for the roberta-large-mnli model), coloured according to listed feature labels.

5.5.2 Visualisations of Interventions

Secondly, we present the same set of representations after the respective *amnesic* and *mnestic* interventions. Recall that these are essentially projections onto or orthogonal to probe decision boundaries with respect to the mentioned label. In figures 5.8 and 5.9, we can assess to what extent amnesic probing and mnestic probing are respectively attaining their stated goals. Recall that amnesic probing works towards erasing the

distinguishability feature classes, while mnestic probing isolates the subspace which best discriminates between classes of the given feature. In figures 5.8 (a) and 5.9 (a), we see that there is a kind of twisting behaviour that may be interpreted as steps in the direction of discouraging the linear separation of classes. Meanwhile, the mnestic results in figures 5.8 (b) and 5.9 (b) present a subspace which is much more visibly segregated between classes. Keep in mind, however, that these are both complementary projections of the same representations in figure 5.7, so neither case is introducing fresh information.

(a) Representations after amnesic intervention with respect to the context monotonicity feature

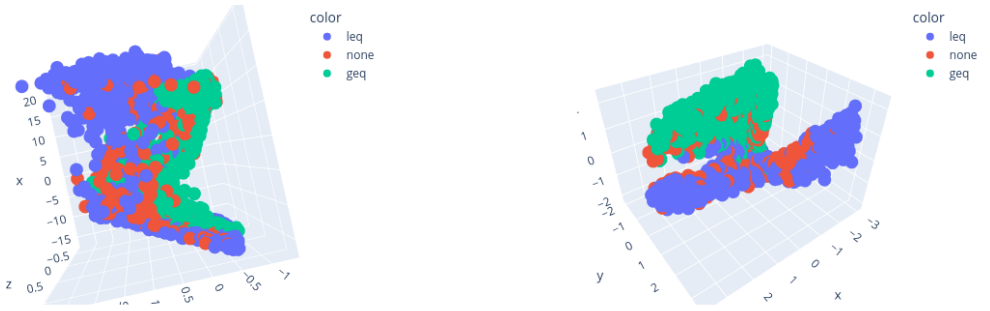
(b) Representations after the mnestic intervention with respect to the context monotonicity feature

Figure 5.8: Projected representations of the `[CLS]` token vectors *after* the a) amnesic and b) mnestic interventions for the NLI-*XY* test set (specifically for the roberta-large-mnli-help model), coloured according to monotonicity feature labels.

5.6 Related Work

The use of probing as an interpretability strategy dates back as far as works such as Alain and Bengio [102] and [112], but a core set of work on the detailed development of the methodology includes Belinkov and Glass [16], Hewitt and Liang [64], Pimentel, Valvoda, Hall Maudslay, *et al.* [69], and Voita and Titov [73]. For a full survey, see Belinkov [32].

The application of probing strategies to natural logic components has been explored in Rozanova, Ferreira, Valentino, *et al.* [108] and Geiger, Richardson, and Potts [84]. In Rozanova, Ferreira, Valentino, *et al.* [108], probing experiments have proven effective



(a) Representations after amnesic intervention with respect to the insertion relation feature

(b) Representations after the mnestic intervention with respect to the insertion relation feature

Figure 5.9: Projected representations of the `[CLS]` token vectors *after* the a) amnesic and b) mnestic interventions for the NLI-*XY* test set (specifically for the roberta-large-mnli-help model), coloured according to insertion relation feature labels.

in detecting the presence or absence of features such as *context monotonicity* and *phrase-pair relations* in the internal representations of NLI models.

Regarding interventions as interpretability tools for machine learning classifiers, there are two broad categories: those that modify the raw input (such as image or text) in a controlled way, and those that modify the hidden/latent vector representations of the data at various stages of the models' input processing. While input-level interventions are more common as they are usually easier to control and are strongly interpretable, they don't allow us to explore and conjecture about exact high-level representational mechanisms in the latent space. We tabulate a few relevant interventional interpretability methods in table 5.2. Note in particular the variation in the *generation* step for the intervened input; some use generative modelling for counterfactual examples, while we use cheaper linear probes.

The only other work in which interventional methods have been applied to natural logic is Geiger, Lu, Icard, *et al.* [81]: a similar problem setting is considered, but at a finer granularity. Our work focuses more on the summarised abstract notion of context monotonicity as a single feature, rather than the intermediate tree nodes that determine its final monotonicity profile. The interventions used in this work are vector *interchange* interventions; partial representations from transformed inputs are used, as opposed to direct manipulations of the encoded vectors.

	Intervention	Tested Effect	Feature Characterisation	Requires Intermediate Labels	Intervention Linked to Concept Interpretation	Domain
Amnesic Probing / INLP [1]	Debiasing / Feature Removal	Downstream Classifier Accuracy	Linear Classifier	Yes	No	Language Modelling
CausaLM: Causal Model Explanation Through Counterfactual Language Models [113]	Re-Training Model Copy For Counterfactual Representation	Text representation-based individual treatment effect (TReITE)	Retrained Base Model	Yes	Yes	Sentiment Analysis
Explaining Classifiers with Causal Concept Effect [114]	Generative Modeling	Average Causal Effect Measure	VAE	Yes	Yes	Vision
Concept Activation Vectors (TCAV) [115]	Value Shift in Vector Direction	Custom Gradient Sensitivity Measure	Linear Classifier	Yes	Yes	Vision
Latent Space Explanation by Intervention [116]	VAE Input Discretization and Reconstruction	Reconstruction Quality	VAE	No	Qualitative Judgement (Vision Only)	Vision
Meaningfully Debugging Model Mistakes Using Conceptual Counterfactuals [117]	Weighted Combination of Concept Vectors	Difference Between Concept Addition and Removal Effect	Linear Classifier	Yes	Yes	Vision

Table 5.2: Related Work on Latent Concept Interventions

5.7 Conclusion and Future Work

In this chapter, we address **RQ 6** (*What can structural interventional methods tell us about the usefulness of the identified representations for the NLI task?*) and **RQ 7** (*How can we devise an alternative interventional interpretability method that is still informative in the high-dimensional situations where amnesic probing fails?*) by carrying out new and existing vector-level interventions to investigate the effect of our semantic features of interest on NLI classification. We perform *amnesic* probing (which removes features as directed by learned probes) and introduce the *mnestic* probing variation (which forgets all dimensions *except* the probe-selected ones). Furthermore, we delve into the limitations of these methods and outline pitfalls that have been obscuring the effectivity of such studies.

In the initial application *amnesic probing* we discover some curious limitations in our high-dimensional settings where there are few label classes (and consequently fewer dimension modified), even if these classes are initially able to be detected with high accuracy by linear probes and amnesic probing shows a strong decrease in probing

scores.

Our results point out that it is misguided to conclude that a given feature is not used when post-amnesic-intervention downstream performance fails to drop, especially in our example amnesic probing studies of a) the gold donwstream feature label and b) the composite of two labels that jointly determine the entailment label. For the gold label "feature" this is especially nonsensical. We hypothesise that due to the low number of feature classes, the effective intervention is too low-rank to meaningfully affect the representations, which may have a lot of redundancy to be exploited by the final classifier. As well-suited as the amnesic probing paradigm is for our setting in theory, its inefficacy leaves a gap for a more informative probing-based interpretability method.

Our introduced *mnestic* variation of the interventional methodology yields much cleaner insights, once again demonstrating the difference between model capabilities before and after the HELP improvement strategy. We now back up previous observations with an *interventional* observation which suggests that both monotonicity and relations are useful for HELP-improved models, while previous state-of-the-art models show a strong use of concept relations but not so much of the monotonicity feature.

It remains to be checked whether high performance in the random control directions corresponds to strong alignment with these probe-selected directions: we propose a potential future analysis of the *dot products* with the fixed set of probe-selected dimensions, which indicates a shared directionality measure (0 for orthogonal vectors and 1 for codirectional ones).

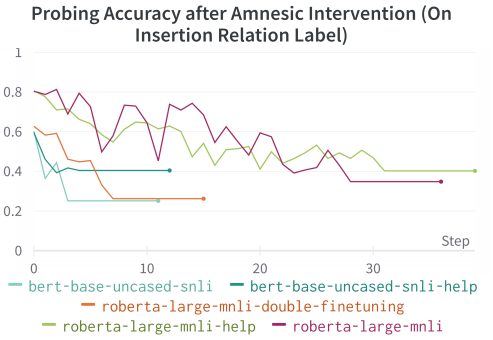
In summary: we have introduced a modification of the amnesic probing paradigm which we call *mnestic* probing which uses the same INLP process but considers the opposite intervention: using the union of projection rowspaces to keep *only* the directions the probes have identified to be modelling the target information. This strategy presents results that are more aligned with theoretical expectations (in the NLI case), possibly because we are now able to make comparisons in a lower rank setting.

5.8 Scoping and Limitations

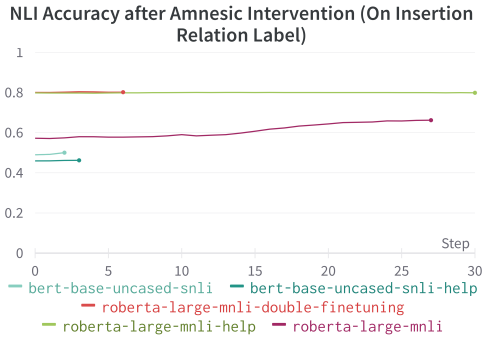
A key limitation of the mnestic probing strategy is that as one reconstructs the original representation one dimension at a time, information content is naturally due to increase: as such, no mnestic probing result can be viewed in isolation, but should be used as a comparative study. Preferably, various randomized selections of linear subspaces with

the same number of dimensions should be included as baselines input representations. Furthermore, we mention two some additional caveats: firstly, the probing strategies used here to identify the informative semantic subspaces in question are always linear; relevant information may be present non-linearly. However, as with amnesic probing, we discount any non-linearly encoded information as the final model classifcation layer is linear and thus cannot exploit this information. Lastly, probing for subspaces which are informative of target auxiliary features may always include correlated features in the resulting subspaces; this must always be taken into account when drawing conclusions from mnestic/amnesic probing.

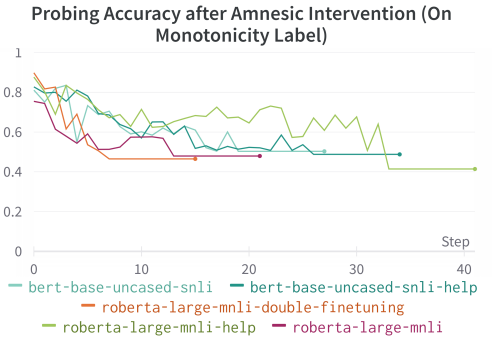
5.9 Expanded Amnesic Intervention Results



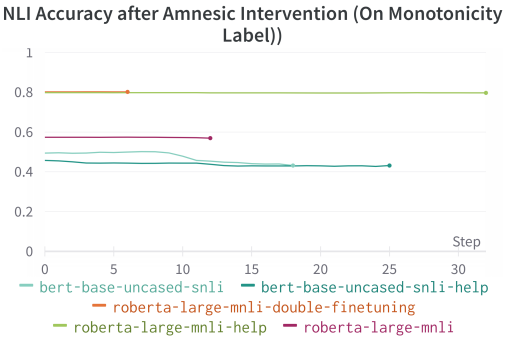
(a) Lexical relation probing performance during iterative amnesic intervention process



(b) Downstream performance on NLI-*XY* after amnesic intervention (removing lexical relation information)

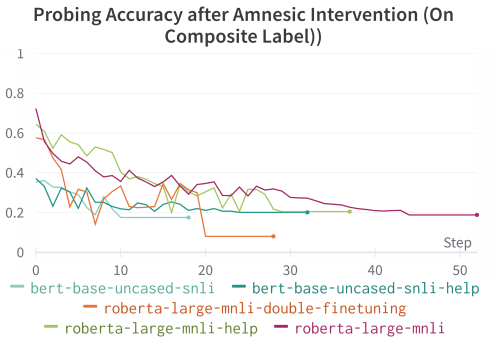


(c) Context monotonicity probing performance during iterative amnesic intervention process

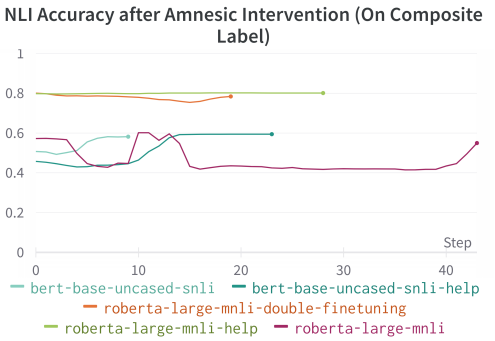


(d) Downstream performance on NLI-*XY* after amnesic intervention (removing context monotonicity information)

Figure 5.10: Single feature label amnesic probing, presenting a) the probing score at each INLP step, and b) the NLI task accuracy when using the representations create at that INLP step.

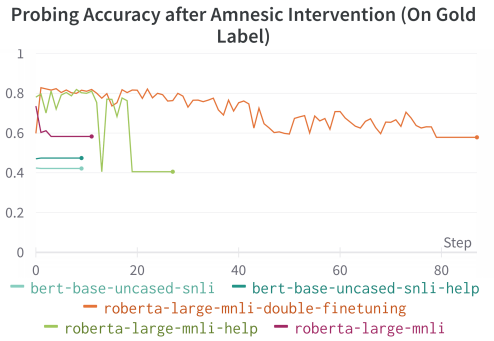


(a) Probing performance on NLI-*XY* after composite label amnesic intervention

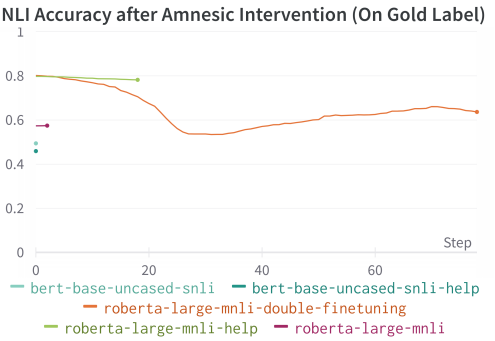


(b) Downstream performance on NLI-*XY* after composite label amnesic intervention

Figure 5.11: Composite feature label amnesic probing, presenting a) the probing score at each INLP step, and b) the NLI task accuracy when using the representations create at that INLP step.



(a) Probing performance On NLI-*XY* after entailment label amnesic intervention.



(b) Downstream performance on NLI-*XY* after entailment label amnesic intervention.

Figure 5.12: Sanity check: entailment gold label amnesic probing, presenting a) the probing score at each INLP step, and b) the NLI task accuracy when using the representations create at that INLP step.

Chapter 6

Causal Effects of Natural Logic Features

This chapter investigates **RQ 8**: "*What can causal effect measures from interventional experiments tell us about NLI models' robustness and sensitivity to different types of intermediate feature changes?*". Drawing from the framework in Stolfo, Jin, Shridhar, *et al.* [2] and adjusting it to our NLI setting, we produce a causal diagram which models the possible influences of context monotonicity, word pair relations and irrelevant textual surface form on NLI model predictions. We arrange examples from the NLI-*XY* dataset into sets of *interventions*, changing or keeping constant the exact selection of variables needed in order to calculate either *total causal effects* of changes which we would like to see changing model predictions (desired total causal effects), or the *direct causal effects* of interventions for which we would not expect or desire model predictions to be affected (undesired direct causal effects). Following Stolfo, Jin, Shridhar, *et al.* [2], we interpret these measures as indicators of *robustness* and *sensitivity* with respect to the semantic variables of interest.

Our central conclusion is that high performance on NLI benchmarks does not coincide with the strong robustness and sensitivity indicatggors. In particular, we bring causal evidence to support previous observations that popular NLI models *fail to respond appropriately to context monotonicity*, and *over-rely on relations between words across the premise and hypothesis*. Furthermore, we bring a new observation that strategies to improve model monotonicity handling (namely, fine-tuning on the HELP dataset) also benefit the treatment of concept relations, even though this was not evident in previous observational studies. Especially, we note that the causal effect indicators in this chapter show that improved models also demonstrate improved *robustness* to

irrelevant insertion pair changes and better *sensitivity* to relevant ones.

6.1 Introduction

There is an abundance of reported cases where high accuracies in NLP tasks can be attributed to simple heuristics and dataset artifacts [7]. As such, when we expect a language model to capture a specific reasoning strategy or correctly use certain semantic features, it has become good practice to perform evaluations that provide a more granular and qualitative view into model behaviour and efficacy. In particular, there is a trend in recent work to incorporate causal measures and *interventional* experimental setups in order to better understand the captured features and reasoning mechanisms of NLP models [2], [34], [35], [81].

In general, it can be hard to pinpoint all the intermediate features and critical representation elements which are guiding the inference behind an NLP task. However, in many cases there are subtasks which have enough semantic/logical regularity to perform stronger analyses and diagnose clear points of failure within larger tasks such as NLI and QA (Question Answering). As soon as we are able to draw a causal diagram which captures a portion of the model's expected reasoning capabillities, we may be guided in the design of interventional experiments which allow us to estimate causal quantities of interest, giving insight into how different aspects of the inputs are used by models.

In this chapter, we look at the structured subset of the NLI task described in section 1.2.1 to investigate the use of two semantic inference features by NLI models: concept inclusion relations and context monotonicity. In the manner of Stolfo, Jin, Shridhar, *et al.* [2], we use these intermediate semantic feature labels to construct *intervention sets* out of NLI examples which allow us to measure certain causal effects. These measures are used to get a sense of models' sensitivity to relevant changes and robustness to irrelevant changes in these features.

Our contributions may be summarised as follows:

- Following Stolfo, Jin, Shridhar, *et al.* [2], we focus on a structured subproblem in NLP (in our case, a natural logic based subtask of NLI) and present a causal diagram which captures both desired and undesired potential reasoning routes which may describe model behaviour.

- We adapt the NLI-*XY* dataset in chapter 4 to a meaningful collection of *intervention sets* which enable the computation of certain causal effects.
- We calculate estimates for undesired direct causal effects and desired total causal effects, which also serve as a quantification of model robustness and sensitivity to our intermediate semantic features of interest.
- We compare a suite of popular NLI models, identifying behavioural weaknesses in high-performing models and behavioural advantages in some worse-performing ones.

We are the first to complement previous observations of NLI models' brittleness with respect to context monotonicity with the evidence of causal effect measures, as well as presenting new insights that over-reliance on lexical relations is consequently also tempered by the same improvement strategies.

6.2 Problem Formulation

6.2.1 A Structured NLI Subtask

As soon as we have a concrete description of how a reasoning problem *should* be treated, we can begin to evaluate how well a model emulates the expected behaviour and whether it is capturing the semantic abstractions at play. As our structured reasoning problem, we once again use the NLI subtask presented in section 1.2.1. We will re-use the NLI-*XY* dataset introduced in chapter 4, but for the convenience of this chapter we change the notation somewhat. We represent an NLI-*XY* example n as a tuple $n = (c, m, w, r, g)$ in which c is the shared natural language context, m is its monotonicity label, w is a pair (w_1, w_2) of nouns/noun phrases which will be inserted into the context (we refer to these as the *inserted word pair* for brevity), r is the concept inclusion relation label for w and g is the entailment gold label arising from m and r as per table 1.4. We denote by $P(Y \mid C = c, W = w)$ the probabilistic output of a trained NLI model with the example n as the NLI input (in particular, the input is the premise–hypothesis pair $(c(w_1), c(w_2))$). The support of the variable Y is the set $\{0, 1\}$: 0 is interpreted as non-entailment and 1 is interpreted as entailment.

As we have chosen a coarse segmentation of the monotonicity reasoning problem, we are able to present a simple causal diagram which illustrates our expectations for the correct reasoning scheme for a fixed class of NLI problems. The diagram in figure 6.1

Variable	Description
G	Gold Label
C	Context
M	Context Monotonicity
W	Inserted Word Pair
R	Word-Pair Relation

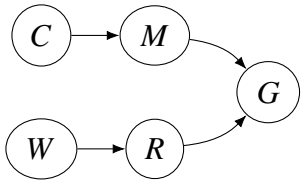


Figure 6.1: Causal diagram for the natural logic subtask.

shows the features on which the gold label is dependent on in the NLI-XY dataset: only the context monotonicity M and the concept pair relation R, which are respectively dependent on the content of the natural language context C and the concept pair / word pair W which is substituted into it. The exact values of the gold label with respect to these features may be referenced in table 1.4.

Naturally, it is always likely that models may fail to follow the described reasoning scheme for these NLI problems. In the next section (6.2.2), we propose a causal diagram which also captures the reasoning possibilities an NLI model may follow, accounting for possible confounding heuristics which may manifest as unwanted direct causal effects.

6.2.2 The Causal Structure of Model Decision-Making

In an ideal situation, a strong NLI model would identify the word-pair relation and the context monotonicity as the abstract variables relevant to the final entailment label. In this case, these features would causally affect the model prediction in the same way they affect the gold label. Realistically, as shown in illuminating studies such as McCoy, Pavlick, and Linzen [7], models identify unexpected biases in the dataset and may end up using accidental correlations output labels, such as the frequency of certain words in a corpus. For example, McCoy and Linzen [31] demonstrate how models can successfully exploit the presence of negation markers to anticipate non-entailment, even when it is not semantically relevant to the output label.

To ensure that the semantic features themselves are taken account into the model's output and not other surface-level confounding variables, one would like to perform interventional studies which alter the value of the target feature but not other confounding

variables. This is, in many cases, not feasible (although attempts are sometimes made to at least perform interventions that only make minimal changes to the textual surface form, as in Kaushik, Hovy, and Lipton [118].)

Stolfo, Jin, Shridhar, *et al.* [2] argue that it is useful to quantify instead the direct impact of irrelevant surface changes (controlling for values of semantic variables of interest) and compare them to *total causal effects* of input-level changes: doing so, we may posit deductions about the flow of information via the semantic variables (or lack thereof). For analyses where there is an attempt to align intermediate variables with explicit internals, see Vig, Gehrmann, Belinkov, *et al.* [34] and Finlayson, Mueller, Gehrmann, *et al.* [35] for a mediation analysis approach, or Geiger, Lu, Icard, *et al.* [81] for an alignment strategy based on causal abstraction theory.

Diagram Specification We follow Stolfo, Jin, Shridhar, *et al.* [2] in the strategy of explicitly modeling the "irrelevant surface form" of the input text portions as variables in the causal diagram. Their setting of *math word problems* is decomposed into two compositional inputs: a question template and two integer arguments. Our setting follows much the same structure: our natural language "context" plays the same role as their "template", but our arguments (an inserted word pair) have an additional layer of complexity as we also model the *relation* between the arguments as an intermediate reasoning variable rather than the values themselves (as such, the structure of their template modeling in their causal diagram is more applicable than the direct way they treat their numerical arguments.)

We present our own causal diagram in figure 6.2. We introduce the textual context C as an input variable, which is further decomposed into more abstract variables: its *monotonicity* M (which directly affects the gold truth G) and the textual surface form S of the context . The other input variable is the word-pair insertion which we will summarise as a single variable W. Once again, W has a potential effect on the model decision through its textual surface form T and via the relation R between the words. The gold truth G is dependent on W and R only. Finally, the outcome variable is the model prediction Y. The paths for which we would like to observe the highest causal effect are the paths to Y from the inputs via M, R and through the gold truth variable G. However, each of S, T, M and R have direct links to the model output Y as well (indicated in red): these are potential direct effects which are *unwanted*. For example, we would not want a model to learn a prediction heuristic based directly on the variable M, such as consistently predicting non-entailment any time a downward monotone context is

Variable	Description
Y	Model Prediction
G	Gold Label
C	Context
M	Context Monotonicity
S	Context Textual Surface Form
W	Inserted Word Pair
R	Word-Pair Relation
T	Word-Pair Textual Surface Form

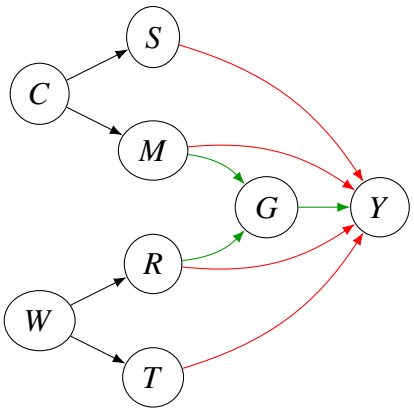


Figure 6.2: Specification of the causal diagram for possible routes of model reasoning for NLI-XY problems. Green edges indicate *desired* causal influence, while red edges indicate *undesired* paths of causal influence via surface-level heuristics.

recognised. Similarly, a direct effect of S or T would look like a heuristic which predicts the entailment label purely based on the presence of words which happened to co-occur with that label in the training data. The key goal of this study is to compare the extent to which models exhibit the high causal effects for the *desired* diagram routes and lower causal effects for the *undesired* routes.

6.3 Estimating the Causal Effects

Given a fixed set N of NLI-XY examples, we define an *intervention* I on N as a set of pairs (n, n') of NLI-XY examples for (one for each $n \in N$), where $n' = (c', m', w', r', g')$ is a second NLI-XY example which represents a modified version of n (in practice, a modification of either c or w). We denote by N' the set of modified NLI-XY examples, so that $I \subseteq N \times N'$.

For any pair $(n, n') \in I$, we define the change-of-prediction indicator

$$CP(n,n') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{if } y = y' \end{cases},$$

where

$$y = \arg\max_{i \in \{0,1\}} P(Y = i \mid C = c, W = w)$$

(namely, the model prediction which assigns the entailment label with the highest predicted probability) and

$$y' = \arg\max_{i \in \{0,1\}} P(Y = i \mid C = c', W = w').$$

Stolfo, Jin, Shridhar, *et al.* [2] refer to the average change-of-prediction quantity for a given intervention I as a *causal effect*. This causal effect quantity is named and interpreted differently depending on the conditions of the intervention: in particular, which variables are changed and which are kept constant throughout the intervention set over which we will take the average.

6.3.1 Interventions for Calculating TCE and DCE

The quantities of interest in Stolfo, Jin, Shridhar, *et al.* [2] are the *total causal effect* (TCE) of interventions on the variables which we would like to see having an effect on the prediction (in our case, C and W) and the *direct causal effect* (DCE) of interventions on the variables which we do *not* wish to unnecessarily impact the model prediction (in our case, T and S).

For a given *source* variable and *target* variable, whether we are measuring a DCE or TCE differs only in the design of the intervention set, which in turn depends on the structure of the causal diagram. For the design of the relevant intervention sets, we follow the strategy in Stolfo, Jin, Shridhar, *et al.* [2], as the upper portion of their causal diagram (concerning the natural language question template, its textual surface form and the implicit math operation) is equivalent to both the upper and lower half of our diagram in figure 6.2.

In this work, we provide four intervention sets: I_0, I_1, I_2, I_3, each corresponding to

the quantities TCE (C on Y), TCE (W on Y), DCE ($T \rightarrow Y$) and DCE ($S \rightarrow Y$) respectively [1]. We stick to their nomenclature of total causal effect (TCE) and direct causal effect (DCE), but define the quantities in the way that they are concretely calculated (in both our experiments and in Stolfo, Jin, Shridhar, *et al.* [2]): as an *estimate* of the causal effect quantity, which they present as an expected value of the change-of-prediction indicator.

(Desired) Total Causal Effects

1. We estimate the total causal effect of the context C on the model prediction Y by constructing an intervention set I_0 as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different context c' from the NLI-XY dataset which should result in a changed prediction, while keeping the inserted word pair w constant.

   In summary, every $(n, n') \in I_0$ satisfies

   $$(c \neq c', m \neq m', w = w', r = r', g \neq g').$$

   We then calculate:

   $$\text{TCE}(C \text{ on } Y) = \frac{1}{|I_0|} \sum_{(n,n') \in I_0} CP(n, n')$$

2. Secondly, we estimate the total causal effect of the inserted word pair W on the model prediction Y by constructing an intervention set I_1 as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different inserted word pair w' from the NLI-XY dataset which should result in a changed prediction, while keeping the shared context c constant.

   In summary, every $(n, n') \in I_1$ satisfies

   $$(c = c', m = m', w \neq w', r \neq r', g \neq g').$$

[1] To be consistent with the notation in Stolfo, Jin, Shridhar, *et al.* [2], we will stylize these quantities as (for example) TCE(C on Y) and DCE($S \rightarrow Y$), where the arrow emphasizes that the quantity is specific to a direct path in the causal diagram (passing through no intermediate variables).

We then calculate:

$$\text{TCE}(W \text{ on } Y) = \frac{1}{|I_1|} \sum_{(n,n') \in I_1} CP(n,n')$$

Following Stolfo, Jin, Shridhar, *et al.* [2], we interpret this quantity as a measure of model *sensitivity* to relevant context (respectively, inserted word pair) changes. As it quantifies how often the prediction changes when it should, we would like to see this value being as close to 1 as possible.

(Undesired) Direct Causal Effects The total causal effect does not distinguish whether this effect is mediated through the preferred causal route (for example, via context's monotonicity) or through a model heuristic based on the textual surface form: it is taking into account all possible routes of influence. The key suggestion in Stolfo, Jin, Shridhar, *et al.* [2] is that even though we have no feasible intervention strategies which allow us to calculate the causal effect of the intermediate variables M and R on Y as mediated through the gold label G (the effect of greatest interest to us), we may yield some insight into their causal influence by comparing the relevant TCE to the unwanted *direct causal effect* DCE ($S \rightarrow Y$) (respectively, DCE ($T \rightarrow Y$)).

1. To estimate the direct causal effect of the textual surface form S of the context C which is irrelevant to the context monotonicity M, we construct an intervention set I_2 as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different context c' from the NLI-XY dataset while conditioning on the monotonicity (specifically, c' is chosen so that its monotonicity attribute m' is the same as that of c). The word pair w' (and therefore its relation r') are kept the same as in n, so the prediction is expected *not* to change. In summary, every $(n,n') \in I_2$ satisfies

$$(c \neq c', m = m', w = w', r = r', g = g').$$

We then calculate:

$$\text{DCE}(S \rightarrow Y) = \frac{1}{|I_2|} \sum_{(n,n') \in I_2} CP(n,n')$$

2. To estimate the direct causal effect of the textual surface form T of the inserted word pair W which is irrelevant to the word pair relation R, we construct an

intervention set I_3 as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different inserted word pair w' from the NLI-XY dataset while conditioning on the word pair relation (specifically, w' is chosen so that its relation attribute r' is the same as that of w). The context c' (and therefore its monotonicity m') are kept the same as in n, so the prediction is expected *not* to change. In summary, every $(n, n') \in I_3$ satisfies

$$(c = c', m = m', w \neq w', r = r', g = g').$$

We then calculate:

$$\text{DCE}(T \rightarrow Y) = \frac{1}{|I_3|} \sum_{(n,n') \in I_3} CP(n, n')$$

Once again following Stolfo, Jin, Shridhar, *et al.* [2], we interpret this quantity as a measure of model *robustness* to irrelevant context (respectively, inserted word pair) changes. As it quantifies how often the prediction changes in cases when it *shouldn't*, we would like to see this value being as close to 0 as possible.

We present examples and dataset statistics for the intervention sets in the next section, along with the summary of the intervention schema in table 6.1.

6.4 Experimental Setup

6.4.1 Data and Interventions

Intervention Set	Target Measure	C	W	M	R	G	Interventions in Dataset
I_0	TCE $(C \rightarrow Y)$	$\neq$	$=$	$\neq$	$=$	$\neq$	14270
I_1	TCE $(W \rightarrow Y)$	$=$	$\neq$	$=$	$\neq$	$\neq$	22640
I_2	DCE $(S \rightarrow Y)$	$\neq$	$=$	$=$	$=$	$=$	20910
I_3	DCE $(T \rightarrow Y)$	$=$	$\neq$	$=$	$=$	$=$	25960

Table 6.1: Intervention schema and dataset statistics: which variables are held constant and which are changed in the construction of intervention sets for the calculation of the indicated effects.

We use the NLI-XY evaluation dataset to construct intervention pairs (n, n') by using a sampling/filtering strategy as in [2] according to the intervention schema in table 6.1, and as described in section 6.3. In particular, for constructing *context* interventions, we

sample a seed set of 400 NLI-*XY* premise/hypothesis pairs. This is the *pre-intervention* NLI example. For each, we fix the insertion pair and filter through the NLI-*XY* dataset for all examples with the shared insertion pair but different context, conditioned as necessary on the properties of the other variables as in the intervention schema. For insertion pairs, we do the opposite. The number of interventions we produce in this way for our experiments are reflected in the last column of table 6.1

In summary, the changes are context replacements and related word-pair replacements; we provide text-level examples in tables 6.2 and 6.3 .

Intervention Set	Target Quantity	Intervention Step	Premise	Hypothesis	M	R	G
I_1	TCE(W on Y)	Before	There's a cat on the pc.	There's a cat on the machine.	$\uparrow$	$\sqsubseteq$	Entailment
		After	There's a cat on the tree.	There's a cat on the fruit tree.	$\uparrow$	$\sqsupseteq$	Non-Entailment
I_3	DCE($T \rightarrow Y$)	Before	There are no students yet.	There are no first-year students yet.	$\downarrow$	$\sqsupseteq$	Entailment
		After	There are no people yet.	There are no women yet.	$\downarrow$	$\sqsupseteq$	Entailment

Table 6.2: Example word-pair insertion interventions for determining the total causal effect of label-relevant word-pair changes and the direct causal effect of label-irrelevant word-pair changes.

Intervention Set	Target Quantity	Intervention Step	Premise	Hypothesis	M	R	G
I_0	TCE(C on Y)	Before	You can't live without fruit .	You can't live without strawberries .	$\uparrow$	$\sqsupseteq$	Non-Entailment
		After	All fruit study english.	All strawberries study English.	$\downarrow$	$\sqsupseteq$	Entailment
I_2	DCE($S \rightarrow Y$)	Before	He has no interest in seafood .	He has no interest in oysters .	$\downarrow$	$\sqsupseteq$	Entailment
		After	I don't want to argue about this in front of seafood .	I don't want to argue about this in front of oysters .	$\downarrow$	$\sqsupseteq$	Entailment

Table 6.3: Example context interventions for determining the total causal effect of label-relevant context changes and the direct causal effect of label-irrelevant context changes.

6.4.2 Model Choice and Benchmark Comparison

We include the following models [2] in our study:

- The models evaluated in NLI-*XY* paper [108], namely roberta-large-mnli, facebook/bart-large-mnli, bert-base-uncased-snli and their counterparts fine-tuned on the HELP dataset [14]

[2] All pretrained models are from the Huggingface *transformers* library ([96]), except for infobert and the pretrained model counterparts fine-tuned on HELP: their sources are linked in the README of the accompanying code.

- The infobert model, which is trained on three benchmark training sets of interest: MNLI [18], SNLI [19] and ANLI [119] (currently at the top of the leaderboard for the adversarial ANLI test set, as of January 2023)

- Another roberta-large checkpoint, also trained on all three benchmark NLI training training sets (as well as FEVER-NLI [120]).

We report their scores on the mentioned benchmark datasets alongside the relevant total and direct causal effects we are interested in.

Note that as the HELP dataset is a two-class entailment dataset (as opposed to datasets like MNLI, which are three-class), we cannot directly compare existing reported scores. As such, we adapt the three-class scores to a two-class score by grouping two of the three-class labels ("contradiction" and "neutral") into the two-class umbrella label "non-entailment". For all models, we report both the three-class and adapted two-class accuracy scores on the benchmark datasets.

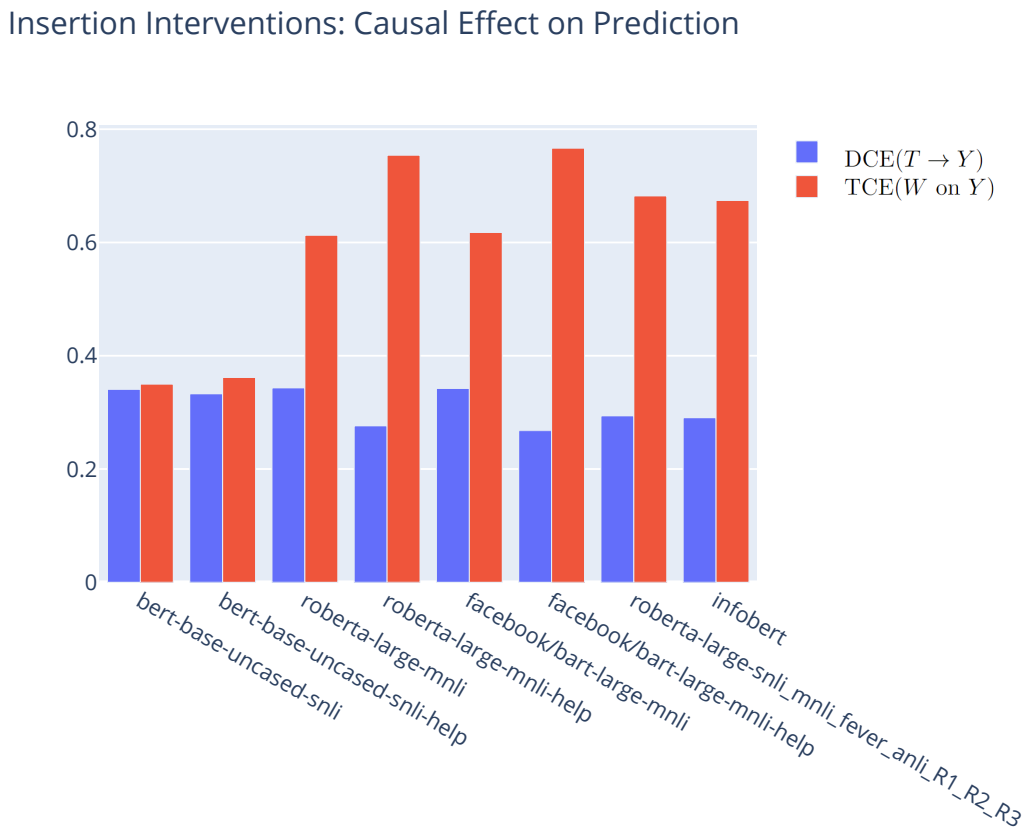
6.5 Results and Discussion

We examine and compare the results for the models listed in 6.4.2. We first look at the word-pair insertion intervention experiments in 6.5.1, then the context intervention experiments in 6.5.2 and finally present a categorical overview of these results in section 6.5.3, contextualised by benchmark scores.

6.5.1 Causal Effect of Inserted Word Pairs

The results for the substituted word-pair intervention experiment are reported in figure 6.3. The most desireable outcome is a DCE($S \rightarrow Y$) which is *as low as possible* in combination with a TCE(C on Y) which is *as high as possible*. The lower this DCE, the higher the model robustness is to *irrelevant context surface form* changes. On the other hand, the higher the specified TCE, the greater the model's sensitivity to *context changes affecting the gold label*.

The largest delta between these two quantities can be seen in the roberta-large-mnli-help and facebook-bart-large-mnli-help models. This is important to note: the HELP dataset [14] is explicitly designed to bolster model success on natural logic problems, but until now there has been little to no evidence that it improves the treatment of word-pair relations. In particular, the internal probing results in [108] show that



Model	DCE($T \rightarrow Y$)	TCE($W \rightarrow Y$)	TCE/DCE Ratio	Delta
bert-base-uncased-snli	0.341	0.350	1.027	0.009
bert-base-uncased-snli-help	0.332	0.361	1.087	0.029
roberta-large-mnli	0.343	0.613	1.785	0.269
roberta-large-mnli-help	0.276	0.754	2.730	0.478
facebook/bart-large-mnli	0.342	0.618	1.805	0.275
facebook/bart-large-mnli-help	0.268	0.766	2.863	0.499
roberta-large-snli_mnli_fever_anli_R1_R2_R3	0.294	0.682	2.321	0.388
infobert	0.291	0.674	2.320	0.384

Figure 6.3: Results for insertion interventions.

probing performance for the intermediate word-pair relation label decreases slightly for roberta-large-mnli after fine-tuning on HELP; as such, it was thought that the HELP improvements on natural logic could solely be attributed to improved context monotonicity treatment. Now, however, we observe distinct improvements in robustness to irrelevant word-pair insertion changes and sensitivity to relevant ones.

More generally, the work in Rozanova, Ferreira, Valentino, *et al.* [108] does indicate

that the large MNLI-based models are already very successful in distinguishing the relation between substituted words. The word-pair relation label has a high *probing* result for all of these models, as well as strong signs of systematicity in their error analysis. This is in line with our observations of relatively large deltas between the DCE and TCE here, compared to the smaller BERT-based models.

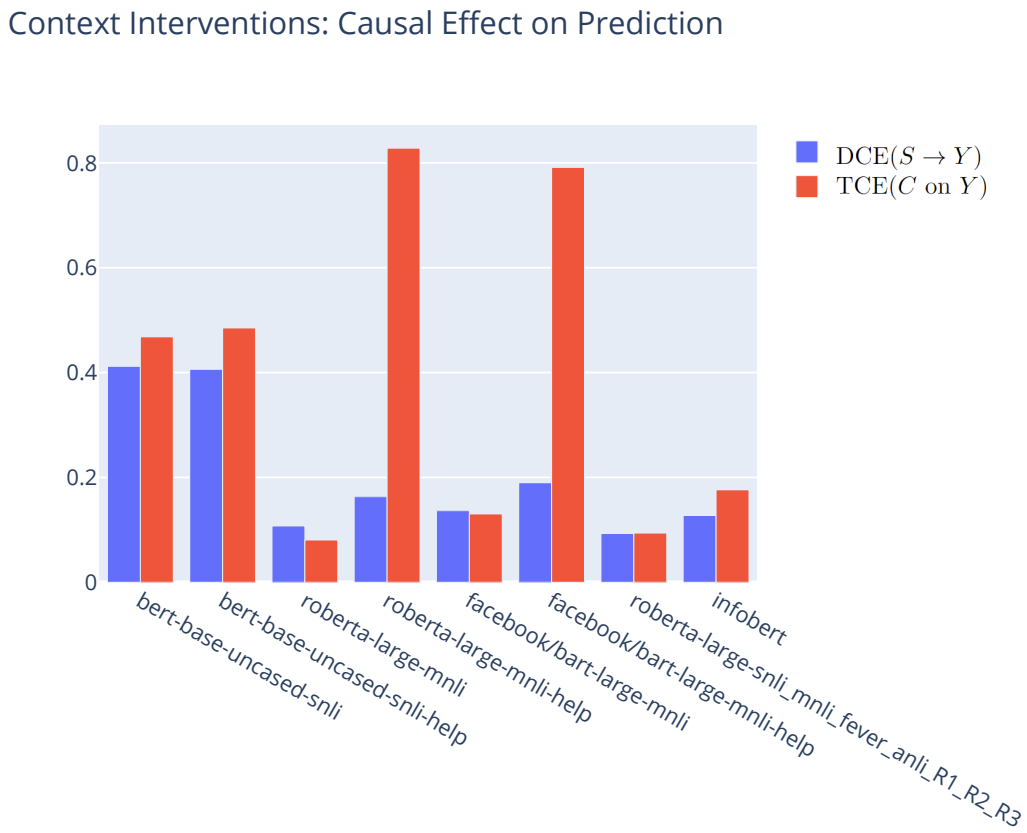
6.5.2 Causal Effect of Contexts

The results for the context intervention experiments are reported in figure 6.4. The most desireable outcome is a DCE($S \rightarrow Y$) which is *as low as possible* in combination with a TCE(C on Y) which is *as high as possible*.

For context interventions, we start to see major distinctions in the sensitivity of models to important context changes – especially the effect of the HELP fine-tuning dataset in increasing model reasoning with respect to context structure. In line with previous behavioural findings in Yanaka, Mineshima, Bekki, *et al.* [13], [14], Geiger, Richardson, and Potts [84], Rozanova, Ferreira, Valentino, *et al.* [108], and Richardson, Hu, Moss, *et al.* [121] and all the way back to Wang, Singh, Michael, *et al.* [6], which observe systematic failure of large language models in downward monotone contexts, we notice that all of the models trained only on the large benchmarks sets fail to correctly change their prediction when a context change requires it to do so (as indicated by the low TCE score).

In Yanaka, Mineshima, Bekki, *et al.* [14], Rozanova, Ferreira, Valentino, *et al.* [108] and Rozanova, Ferreira, Thayaparan, *et al.* [110], the positive effect of the HELP dataset is already evident, but here we may also compare it to roberta-large-mnli tuned on many additional training sets, precluding the possibility that its helpfulness can be attributed only to a "larger amount of training data".

We note that although the situation of the TCE/DCE ratio for roberta-large-mnli being less than one may seem peculiar, it is important to keep in mind that the intervention sets used for estimating these quantities are sampled independently so some margin of error is warranted. As in Stolfo, Jin, Shridhar, *et al.* [2], we interpret this result to simply mean that the causal influence is comparable whether we are affecting the ground truth result (as in the TCE($C \rightarrow Y$) case) or not (as in the DCE($S \rightarrow Y$) case).



Model	DCE($S \rightarrow Y$)	TCE($C \rightarrow Y$)	TCE/DCE Ratio	Delta
bert-base-uncased-snli	0.412	0.468	1.136	0.0563
bert-base-uncased-snli-help	0.406	0.485	1.194	0.079
roberta-large-mnli	0.107	0.081	0.751	-0.027
roberta-large-mnli-help	0.163	0.828	5.070	0.665
facebook/bart-large-mnli	0.136	0.130	0.954	-0.006
facebook/bart-large-mnli-help	0.189	0.791	4.167	0.601
roberta-large-snli_mnli_fever_anli_R1_R2_R3	0.093	0.093	1.008	0.001
infobert	0.127	0.176	1.385	0.049

Figure 6.4: Results for context interventions.

6.5.3 Benchmark Scores and Causal Effects

A summary of the performance of all models on popular benchmarks alongside a categorical breakdown of robustness and sensitivity is presented in table 6.4. The robustness/sensitivity categories are a qualitative assessment, identifying the *lowest* and *highest* scores within a category, and categorising other models correspondingly as *low*,

Model	NLI Benchmark Evaluation (2 Class Accuracy)						Context Changes		Inserted Word-Pair Changes	
	SNLI	MNLI-M	MNLI-MM	ANLI-R1	ANLI-R2	ANLI-R3	Robustness	Sensitivity	Robustness	Sensitivity
bert-base-uncased-snli	0.766	0.620	0.623	0.567	0.596	0.580	Mid	Mid	Mid	Low
bert-base-uncased-snli-help	0.757	0.627	0.626	0.505	0.508	0.546	Mid	Mid	Mid	Low
facebook/bart-large-mnli	0.935	0.940	0.939	0.596	0.563	0.593	High	Low	Mid	Mid
facebook/bart-large-mnli-help	0.727	0.802	0.795	0.538	0.489	0.528	Mid/High	**Highest**	**Highest**	**Highest**
roberta-large-mnli	0.931	0.941	0.940	0.614	0.529	0.5325	**Highest**	Lowest	Mid	Mid
roberta-large-mnli-help	0.738	0.668	0.656	0.565	0.554	0.574	High	**Highest**	**Highest**	**Highest**
roberta-large-snli_mnli_fever_anli	0.949	0.936	0.939	0.810	0.659	0.666	**Highest**	Lowest	Mid	Mid/High
infobert	**0.950**	**0.943**	**0.941**	**0.837**	**0.682**	**0.683**	High	Low	Mid	Mid/High

Table 6.4: Overall two-class accuracy on original NLI benchmarks and qualitative comparison against the performed causal intervention analysis. The accuracy is not necessarily predictive of the performances achieved using a systematic causal inspection.

mid or *high* performers for the given categories. The sensitivity property is tied to the desired total causal effect, while the robustness property is tied to the undesired direct causal effect (note in particular that the latter is judged as *inversely proportional:* the model with the lowest given DCE is judged the "highest" in terms of robustness).

The key observation is that the models which achieve the highest performance on benchmarks may be far from the best performers with respect to our quantitative markers of strong reliance of important causal features. In particular, models such as infobert are outperformed in our behavioural causal effect analyses by weaker models that are fine-tuned on a relatively small helper dataset such as HELP. It is important to note that such changes coincide with drops in benchmarks performance too, but any model interventions that discourage the exploitation of heuristics (evident from a lower DCE for surface form features) may have that effect.

6.6 Related Work

Causal modelling has appeared in NLP works in various forms, such as the investigations of the causal influence of data statistics [122] and mediation analyses [34], [35] which link intermediate linguistic/semantic features to model internals. Stolfo, Jin, Shridhar, *et al.* [2], our core reference, appears to be the first to use explicitly causal effect measures as indicators of sensitivity and robustness (for some non-causal approaches to measuring model robustness in NLP, we point to [123] and [124]). For a fuller summary of the use of causality in NLP, see the survey by Feder, Keith, Manzoor, *et al.* [125].

Specific to natural logic, works with causal approaches include Geiger, Richardson, and Potts [15] (which perform interchange interventions at a token representation level), Geiger, Lu, Icard, *et al.* [81] (where an ambitious causal abstraction experiment attempts to align model internals with candidate causal models) and the works of Geiger,

Richardson, and Potts [15] and [126], (where attempts are made to build a prescribed causal structure into models themselves). In particular, [126] create a "causal proxy model" which becomes the basis for a new explainable predictor designed to replace the original neural network.

6.7 Conclusion

In this chapter, we addressed **RQ 8**: (*What can causal effect measures from interventional experiments tell us about NLI models' robustness and sensitivity to different types of intermediate feature changes?*) by drawing from the framework in Stolfo, Jin, Shridhar, *et al.* [2] and adjusting it to our NLI setting. To this end, we produce a causal diagram which models context monotonicity and word pair relations and the possible influences of irrelevant textual surface form. We arrange examples from the NLI-*XY* dataset into sets of *interventions*, changing or keeping constant the exact selection of variables needed in order to calculate either desired *total causal effects* or undesired *direct causal effects*. Following Stolfo, Jin, Shridhar, *et al.* [2], we interpret these measures as indicators of *robustness* and *sensitivity* with respect to the semantic variables of interest.

The results here strongly bolster the fact that similar benchmark accuracy scores may be observed for models that exhibit very different behaviour, especially with respect to specific semantic reasoning patterns and higher-level properties such as robustness/sensitivity with respect to target features. In this chapter, we have been able to explicitly observe previously suspected biases in certain large NLI models. For example, previous observations [13], [108] that roberta-large-mnli is biased in favour of assuming upward-monotone contexts, ignoring the effects of things like negation markers, agrees with our observations that it exhibits poor context sensitivity (a low TCE influence of contexts which should be changing the output label). Furthermore, the causal flavour of the study adds a complementary narrative to works that investigate model internals via probing [108] and observe the presence/absence of intermediate semantic features in model representations. Instead of merely suggesting that these features are captured, we are able to gain insight into their causal influence via connected causal effect estimates. The causal measures presented here show us that even the highest-performing models systematically show a failure to adapt their predictions to changing context structure, suggesting an over-reliance on word relations across the premise and hypothesis.

Lastly, we bring a new observation that strategies to improve model monotonicity handling (namely, fine-tuning on the HELP dataset) also benefit the treatment concept relations, although this was not evident in previous observational studies. Especially, we note that the causal effect indicators in this chapter show that improved models also demonstrate improved *robustness* to irrelevant concept pair changes and better *sensitivity* to relevant ones.

6.8 Scoping and Limitations

Pretrained NLI models often differ in their labelling schemes (among themselves, and from dataset labelling schemes). We have to the best of our ability attempted to cross-check the correctness of label configurations used, but there is always the possibility in this case, and we encourage reproduction and checking of the results before reporting them in any external works.

The causal modeling in section 6.2.2 draws heavily on the work in Stolfo, Jin, Shridhar, *et al.* [2], and we place some trust in the authors' choice of names for the causal effect measures with respect to the chosen intervention schemes, but this does not affect our conclusions with regards to the interpretation of the measures provided as indicators of robustness and sensitivity. We also point out that the causal effect measures presented here are sample-specific estimates, which could potentially be improved upon with even larger samples of data.

Chapter 7

Conclusion

7.1 Summary and Conclusion

In this thesis, we have considered the overarching research question **RQ 0**: *"How well do existing NLI models perform natural logic deductions, and to what extent are they implicitly modelling context monotonicity and concept inclusion relations to do so in a systematic way?"*. After highlighting the previously-observed insufficiencies in state-of-the-art NLI models' ability to reason in downward monotone contexts and observing the improvements afforded by fine-tuning strategies in chapter 3 (addressing **RQ 1** [1]), we turned to model interpretability methods to design experiments which would help us determine the extent to which the intermediate reasoning features are captured and used by the models we compare.

Firstly, we have summarised the landscape of relevant interpretability methods which are being widely used for the study of NLP models in chapter 2 (in an attempt to address **RQ 2** [2]). The interpretability studies we have carried out in this work are enabled by the introduction of compositional NLI-*XY* dataset in chapter 4 (in response to **RQ 3** [3]), in which each NLI example is enriched with the intermediate feature labels for context monotonicity and the concept inclusion relation which jointly give rise to the gold label. Our experimental findings all support the previously-stated hypothesis [13] that the models trained only on only benchmark NLI datasets (such as MNLI and SNLI)

[1] How much does fine-tuning on the HELP dataset improve NLI models' performance on existing natural logic evaluation datasets? Would a secondary transfer-learning task based on the prediction of context monotonicity result in an improvement in overall evaluation scores?

[2] Which interpretability methods are best-suited for our interest in detecting emergent intermediate features?

[3] How can we construct a natural logic dataset that is suitable for both targeted evaluation and interpretability?

have failed to develop a useful notion of context monotonicity. We find that the same models further fine-tuned on the HELP dataset display a variety of qualitative indicators that both context monotonicity and concept inclusion relations are strongly modelled and relied on for predictions. These qualitative indicators include the results of both *observational* interpretability experiments and *interventional* interpretability experiments. The observational interpretablity approaches include the probing, visualisation and error analysis experiments in chapter 4 (in response to the research questions **RQ 4** [4] and **RQ 5** [5]). The interventional interpretability work includes the interventional probing experiments in chapter 5 (addressing **RQ 6** [6] and **RQ 7** [7]) and the estimation of causal effects in chapter 6 (in response to **RQ 8** [8]).

Instances where these interpretability-based evaluations demonstrate strong indicators that crucial intermediate features are captured provide greater confidence that model reasoning patterns are following theoretically-expected strategies. For example, the roberta-large-mnli-help model demonstrates stronger probing scores for the context monotonicity feature, visually discernible clusters for upward and downward monotone contexts in its vector representations and a stronger robustness to irrelevant context interventions than the roberta-large-mnli model. As such, we are much more likely to deduce that roberta-large-mnli-help takes the context monotonicity factor into account rather than over-relying on the concept inclusion relation value, as roberta-large-mnli seems to do.

Aside from the observations specific to NLI models and their treatment of intermediate natural logic features, a core contribution of this thesis has been the introduction of the *mnestic probing* method in response to limitations of the *amnesic probing* method that we have observed. The application of amnesic probing in our experimental setting yielded some unexpected results, such as the lack of effect on the NLI performance score of amnesic interventions which "remove" crucial information, including the gold label itself. Hypothesizing that the issue relates to low rank amnesic transformations with respect to the high dimensionality of the vector representations, we introduced an alternative way to use the outputs of the INLP process carried out in the amnesic probing

[4] Are the intermediate features of context monotonicity and concept inclusion relations emergent in the internal representations of NLI models which perform better at natural logic tasks?

[5] Which features are responsible for errors in poorer-performing models?

[6] "What can structural interventional methods tell us about the usefulness of the identified representations for the NLI task?"

[7] "How can we devise an alternative interventional interpretability method that is still informative in the high-dimensional situations where amnesic probing fails?"

[8] What can causal effect measures from interventional experiments tell us about NLI models' robustness and sensitivity to different types of intermediate feature changes?

strategy: the *mnestic* probing variation. As the outputs of the mnestic interventions were of a lower rank, we were able to make more useful comparative observations that were more in line with our expectations (and with the probing observations in other chapters). The mnestic probing methodology can be useful in any other situations where the model representations have a high number of dimensions while the probing task has a low number of target classes.

In summary, a key motivation of this thesis has been that accuracy scores on benchmark datasets and even targeted challenge sets offer us limited information about model reasoning mechanisms and learned features. Given a selection of models trained for the same task, we have been interested in observing more qualitative differences between the representational structures and behavioural patterns of models that achieve high accuracy scores and those that do not. In this work, we have demonstrated how a set of expectations for the abstractions an NLI model would need to learn in order to acquire a successful reasoning approach for a specific task can guide the design of interpretability experiments which allow us to observe how the stronger presence of these features corresponds to better task performance. Even more importantly, we imagine that well-chosen interpretability strategies could potentially bring to light distinctions in models that may perform similarly on a given test set, but will eventually differ in their ability to generalise well outside of the given test set's domain.

7.2 Opportunities for Future Work

We present a few suggestions for possible future directions which expand on our intersection of interpretability and natural logic handling.

Multi-Hop Inferences and Individual Operators How does a given model treat a pair of substitutions, one which occurs in an upward monotone position and another in a downward monotone one? One route of potential expansion for the NLI-*XY* dataset is to create such examples, which are implicitly testing how models can perform multiple monotonicity reasoning steps in a single NLI example. This is akin to identifying the "order of operations" in an arithmetic model, such as the setting explored in Giulianelli, Harding, Mohnert, *et al.* [61]. Alternatively, one can extend this "order-of-operations" perspective to the study of how the monotonicity of individual linguistic operators affects a final entailment classification in which the effect of multiple operators needs to be taken into account.

Causal Mediation Analysis The causal effect analysis in chapter 6 is more behavioural than structural, looking at the effect of modified textual inputs on classification outputs. However, the strategy of *causal mediation analysis* quantifies the extent to which a causal effect is *mediated* by a given structural component, such as a vector representations or model weights. Excellent examples of causal mediation analysis applied to NLP problems include Vig, Gehrmann, Belinkov, *et al.* [34] and Finlayson, Mueller, Gehrmann, *et al.* [35]. With respect to our setting, it would be interesting to examine how well the probe-identified semantic subspaces corresponding to a given feature mediate its causal effect on the prediction.

Layer-Wise and Training Dynamics As we have been concerned with the structure of final model predictions, our experiments all treat the final representation layer of the models in question. However, it could potentially be interesting to investigate how and where the relevant feature information arises across layers, or across fine-tuning steps during training. D. Hupkes

Generalisation D. Hupkes How methodically are models applying systematic reasoning methods to unexpected instatiations? The compositional NLI-*XY* dataset may include unexpected examples such as "I swallowed a chair" entails "I swallowed furniture". It would be interesting to investigate how strongly model errors correlate to low-probability substitutions, and perhaps this can be achieved with OOD detection metrics (such as energy-based OOD detection [127]) or simply the masked language modelling probability of a token in a given context (this would be harder to apply to multi-token insertions, however). This question links to the idea of *OOD generalisation* in models, which may be a relevant direction to branch towards. For a relevant recent survey, see Hupkes, Giulianelli, Dankers, *et al.* [128].

Architectural Hypotheses As we are only comparing trained transformer models to *each other* in this work, we are limited in our ability to attribute success or failure of monotonicity modelling to specific architectural innovations. For the most part, our conclusions relate to the quality and structure of the data being trained on. However, it would be interesting to draw comparisons to the representations and representational capacities of earlier architectures: for example, with similar training strategies, would BiLSTMs (which also produce contextual embeddings, and are bidirectional in nature) similarly be able to model a high-level monotonicity feature? Or can we build experimental arguments that transformers are inherently better-equipped for this flavour of

task?

7.3 Longevity of This Work

We see the structural interpretability work here as a useful evaluation framework for any NLI models that have intermediate vector representations. The interpretability experiments provide a deeper view of model internals which can show strong distinctions between models with similar benchmark performance, highlighting differences with respect to the treatment of crucial intermediate features. As models increasingly rely on large crowd-source training datasets rather than high-quality training sets which exhibit a nuanced selection of reasoning phenomena, it is important to instead use expert-curated logic-based datasets for many levels of behavioural and structural evaluation.

Beyond NLI, the interpretability methods explored here provide a broad selection of the kinds of experiments that may be helpful in any setting where one may wish to enquire about the use of task-specific intermediate features for any given model task. Lastly, our introduction of the *mnestic probing* method adds an additional interpretability tool which is relevant in any situation where the iterative nullspace projection method can be applied, especially when amnesic probing may not prove to be informative.

7.4 Ethical Implications

Especially with NLU tasks such as natural language inference, the phrasing of the task being stated as detecting entailment or implication may create a false sense of these models being necessarily logical in nature, which is widely observed to be far from the case [7]. Rigid evaluation of the reasoning strategies and intermediate states of large language models is extremely important for scrutinising their predictions, as blind reliance on model outputs can be troublesome in any real-world applications. However, as much as favourable interpretability indicators may increase confidence in model performance, it must be take into consideration that model behaviour may still be erratic and unreliable for out-of-domain examples.

Bibliography

[1] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg, *Amnesic probing: Behavioral explanation with amnesic counterfactuals*, 2020. eprint: arXiv:2006.00995.

[2] A. Stolfo, Z. Jin, K. Shridhar, B. Schölkopf, and M. Sachan, *A causal framework to quantify the robustness of mathematical reasoning with language models*, 2022. arXiv: 2210.12023 [cs.CL].

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[5] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. arXiv: 1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692.

[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. [Online]. Available: https://aclanthology.org/W18-5446.

[7] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. [Online]. Available: https://aclanthology.org/P19-1334.

[8] M. Glockner, V. Shwartz, and Y. Goldberg, "Breaking NLI systems with sentences that require simple lexical inferences," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 650–655. DOI: 10.18653/v1/P18-2103. [Online]. Available: https://aclanthology.org/P18-2103.

[9] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 107–112. DOI: 10.18653/v1/N18-2017. [Online]. Available: https://aclanthology.org/N18-2017.

[10] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. [Online]. Available: https://aclanthology.org/P19-1452.

[11] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4129–4138. DOI: 10.18653/v1/N19-1419. [Online]. Available: https://www.aclweb.org/anthology/N19-1419.

[12] K. Richardson, H. Hu, L. S. Moss, and A. Sabharwal, "Probing natural language inference models through semantic fragments," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative*

Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 8713–8721. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6397.

[13] H. Yanaka, K. Mineshima, D. Bekki, *et al.*, "Can neural networks understand monotonicity reasoning?" In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 31–40. DOI: 10.18653/v1/W19-4804. [Online]. Available: https://www.aclweb.org/anthology/W19-4804.

[14] H. Yanaka, K. Mineshima, D. Bekki, *et al.*, "HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning," in *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 250–255. DOI: 10.18653/v1/S19-1027. [Online]. Available: https://www.aclweb.org/anthology/S19-1027.

[15] A. Geiger, K. Richardson, and C. Potts, "Neural natural language inference models partially embed theories of lexical entailment and negation," in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Online: Association for Computational Linguistics, Nov. 2020, pp. 163–173. DOI: 10.18653/v1/2020.blackboxnlp-1.16. [Online]. Available: https://aclanthology.org/2020.blackboxnlp-1.16.

[16] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 49–72, 2019. DOI: 10.1162/tacl_a_00254. [Online]. Available: https://aclanthology.org/Q19-1004.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[18] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: http://aclweb.org/anthology/N18-1101.

[19] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. [Online]. Available: https://www.aclweb.org/anthology/D15-1075.

[20] R. Cooper, D. Crouch, J. Van Eijck, *et al.*, "Using the framework," Technical Report LRE 62-051 D-16, The FraCaS Consortium, Tech. Rep., 1996.

[21] V. Sanchez, "Studies on natural logic and categorial grammar," 1991.

[22] J. van Benthem, *Essays in Logical Semantics*. Springer, 1986.

[23] B. MacCartney and C. D. Manning, "Natural logic for textul inference," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague: Association for Computational Linguistics, Jun. 2007, pp. 193–200. [Online]. Available: https://www.aclweb.org/anthology/W07-1431.

[24] H. Hu and L. Moss, "Polarity computations in flexible categorial grammar," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 124–129. DOI: 10.18653/v1/S18-2015. [Online]. Available: https://www.aclweb.org/anthology/S18-2015.

[25] H. Hu, Q. Chen, K. Richardson, A. Mukherjee, L. S. Moss, and S. Kuebler, "MonaLog: A lightweight system for natural language inference based on monotonicity," in *Proceedings of the Society for Computation in Linguistics 2020*, New York, New York: Association for Computational Linguistics, Jan. 2020, pp. 334–344. [Online]. Available: https://aclanthology.org/2020.scil-1.40.

[26] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig, "Stress test evaluation for natural language inference," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2340–2353. [Online]. Available: https://aclanthology.org/C18-1198.

[27] A. Geiger, I. Cases, L. Karttunen, and C. Potts, "Stress-testing neural models of natural language inference with multiply-quantified sentences," *ArXiv*, vol. abs/1810.13033, 2018.

[28] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme, "Hypothesis only baselines in natural language inference," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 180–191. DOI: 10.18653/v1/S18-2023. [Online]. Available: https://aclanthology.org/S18-2023.

[29] Y. Nie, Y. Wang, and M. Bansal, "Analyzing compositionality-sensitivity of nli models," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA: AAAI Press, 2019, ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33016867. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33016867.

[30] P. Jeretic, A. Warstadt, S. Bhooshan, and A. Williams, "Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 8690–8705. DOI: 10.18653/v1/2020.acl-main.768. [Online]. Available: https://aclanthology.org/2020.acl-main.768.

[31] R. T. McCoy and T. Linzen, "Non-entailed subsequences as a challenge for natural language inference," *ArXiv*, vol. abs/1811.12112, 2018.

[32] Y. Belinkov, "Probing classifiers: Promises, shortcomings, and advances," *Computational Linguistics*, vol. 48, no. 1, pp. 207–219, Mar. 2022. DOI: 10.1162/coli_a_00422. [Online]. Available: https://aclanthology.org/2022.cl-1.7.

[33] A. Ravichander, Y. Belinkov, and E. Hovy, "Probing the probing paradigm: Does probing accuracy entail task relevance?" In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, 2021, pp. 3363–3377. DOI: 10.18653/v1/2021.eacl-main.295. [Online]. Available: https://aclanthology.org/2021.eacl-main.295.

[34] J. Vig, S. Gehrmann, Y. Belinkov, *et al.*, "Investigating gender bias in language models using causal mediation analysis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 12 388–12 401. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

[35] M. Finlayson, A. Mueller, S. Gehrmann, S. Shieber, T. Linzen, and Y. Belinkov, "Causal analysis of syntactic agreement mechanisms in neural language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1828–1843. DOI: 10.18653/v1/2021.acl-long.144. [Online]. Available: https://aclanthology.org/2021.acl-long.144.

[36] A. Ettinger, "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 34–48, 2020. DOI: 10.1162/tacl_a_00298. [Online]. Available: https://aclanthology.org/2020.tacl-1.3.

[37] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualising and understanding recurrent networks," *CoRR*, vol. abs/1506.02078, 2015. arXiv: 1506.02078. [Online]. Available: http://arxiv.org/abs/1506.02078.

[38] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualising and understanding neural models in NLP," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 681–691. DOI: 10.18653/v1/N16-1082. [Online]. Available: https://aclanthology.org/N16-1082.

[39] X. Shi, I. Padhi, and K. Knight, "Does string-based neural MT learn source syntax?" In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1526–1534. DOI: 10.18653/v1/D16-1159. [Online]. Available: https://aclanthology.org/D16-1159.

[40] R. Ghaeini, X. Fern, and P. Tadepalli, "Interpreting recurrent and attention-based neural models: A case study on natural language inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4952–4957. DOI: 10.18653/v1/D18-1537. [Online]. Available: https://aclanthology.org/D18-1537.

[41] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html.

[42] S. Serrano and N. A. Smith, "Is attention interpretable?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. [Online]. Available: https://aclanthology.org/P19-1282.

[43] J. Wang, J. Tuyls, E. Wallace, and S. Singh, "Gradient-based analysis of NLP models is manipulable," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 247–258. DOI: 10.18653/v1/2020.findings-emnlp.24. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.24.

[44] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. DOI: 10.18653/v1/D19-1002. [Online]. Available: https://aclanthology.org/D19-1002.

[45] M. L. Leavitt and A. Morcos, "Towards falsifiable interpretability research," *arXiv preprint arXiv:2010.12016*, 2020.

[46] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. [Online]. Available: https://aclanthology.org/2020.acl-main.386.

[47] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural nlp: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, Dec. 2022, ISSN: 0360-0300. DOI: 10.1145/3546577. [Online]. Available: https://doi.org/10.1145/3546577.

[48] Y. Belinkov, S. Gehrmann, and E. Pavlick, "Interpretability and analysis in neural NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Online: Association for Computational Linguistics, Jul. 2020, pp. 1–5. DOI: 10.18653/v1/2020.acl-tutorials.1. [Online]. Available: https://aclanthology.org/2020.acl-tutorials.1.

[49] A. Ravichander, Y. Belinkov, and E. Hovy, "Probing the probing paradigm: Does probing accuracy entail task relevance?" In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 3363–3377. DOI: 10.18653/v1/2021.eacl-main.295. [Online]. Available: https://aclanthology.org/2021.eacl-main.295.

[50] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1499–1509. DOI: 10.18653/v1/D18-1179. [Online]. Available: https://aclanthology.org/D18-1179.

[51] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant, "OLMpics-on what language model pre-training captures," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 743–758, 2020. DOI: 10.1162/tacl_a_00342. [Online]. Available: https://aclanthology.org/2020.tacl-1.48.

[52] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties," in *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2126–2136. DOI: 10.18653/v1/P18-1198. [Online]. Available: https://aclanthology.org/P18-1198.

[53] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, IJCNN 2005, ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2005.06.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005001206.

[54] I. Tenney, P. Xia, B. Chen, *et al.*, "What do you learn from context? probing for sentence structure in contextualized word representations," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SJzSgnRcKX.

[55] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen, "Probing pretrained language models for lexical semantics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 7222–7240. DOI: 10.18653/v1/2020.emnlp-main.586. [Online]. Available: https://aclanthology.org/2020.emnlp-main.586.

[56] T. Linzen, E. Dupoux, and Y. Goldberg, "Assessing the ability of LSTMs to learn syntax-sensitive dependencies," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 521–535, 2016. DOI: 10.1162/tacl_a_00115. [Online]. Available: https://aclanthology.org/Q16-1037.

[57] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg, "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=BJh6Ztuxl.

[58] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. [Online]. Available: https://aclanthology.org/N19-1112.

[59] X. Shi, I. Padhi, and K. Knight, "Does string-based neural MT learn source syntax?" In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1526–1534. DOI: 10.18653/v1/D16-1159. [Online]. Available: https://aclanthology.org/D16-1159.

[60] Y. Belinkov, L. Màrquez, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass, "Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 1–10. [Online]. Available: https://aclanthology.org/I17-1001.

[61] M. Giulianelli, J. Harding, F. Mohnert, D. Hupkes, and W. Zuidema, "Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 240–248. DOI: 10.18653/v1/W18-5426. [Online]. Available: https://aclanthology.org/W18-5426.

[62] K. Zhang and S. Bowman, "Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 359–361. DOI: 10.18653/v1/W18-5448. [Online]. Available: https://aclanthology.org/W18-5448.

[63] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3651–3657. DOI: 10.18653/v1/P19-1356. [Online]. Available: https://aclanthology.org/P19-1356.

[64] J. Hewitt and P. Liang, "Designing and interpreting probes with control tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

Processing (EMNLP-IJCNLP), Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2733–2743. DOI: 10.18653/v1/D19-1275. [Online]. Available: https://www.aclweb.org/anthology/D19-1275.

[65] J. Jumelet, M. Denic, J. Szymanik, D. Hupkes, and S. Steinert-Threlkeld, "Language models use monotonicity to assess NPI licensing," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 4958–4969. DOI: 10.18653/v1/2021.findings-acl.439. [Online]. Available: https://aclanthology.org/2021.findings-acl.439.

[66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceedings of Workshop at ICLR*, vol. 2013, Jan. 2013.

[67] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162.

[68] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. [Online]. Available: https://aclanthology.org/N18-1202.

[69] T. Pimentel, J. Valvoda, R. Hall Maudslay, R. Zmigrod, A. Williams, and R. Cotterell, "Information-theoretic probing for linguistic structure," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4609–4622. DOI: 10.18653/v1/2020.acl-main.420. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.420.

[70] Z. Zhu and F. Rudzicz, "An information theoretic view on selecting linguistic probes," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 9251–9262. DOI: 10.18653/v1/2020.emnlp-main.744. [Online]. Available: https://aclanthology.org/2020.emnlp-main.744.

[71] J. Hewitt, K. Ethayarajh, P. Liang, and C. Manning, "Conditional probing: Measuring usable information beyond a baseline," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1626–1639. DOI: 10.18653/v1/2021.emnlp-main.122. [Online]. Available: https://aclanthology.org/2021.emnlp-main.122.

[72] J. Kunz and M. Kuhlmann, "Where does linguistic information emerge in neural language models? measuring gains and contributions across layers," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4664–4676. [Online]. Available: https://aclanthology.org/2022.coling-1.413.

[73] E. Voita and I. Titov, "Information-theoretic probing with minimum description length," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 183–196. DOI: 10.18653/v1/2020.emnlp-main.14. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.14.

[74] T. Pimentel, N. Saphra, A. Williams, and R. Cotterell, "Pareto probing: Trading off accuracy for complexity," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 3138–3153. DOI: 10.18653/v1/2020.emnlp-main.254. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.254.

[75] T. Pimentel and R. Cotterell, "A bayesian framework for information-theoretic probing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2869–2887. DOI: 10.18653/v1/2021.emnlp-main.229. [Online]. Available: https://aclanthology.org/2021.emnlp-main.229.

[76] J. Kunz and M. Kuhlmann, "Test harder than you train: Probing with extrapolation splits," in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 15–

25. DOI: 10.18653/v1/2021.blackboxnlp-1.2. [Online]. Available: https://aclanthology.org/2021.blackboxnlp-1.2.

[77] J. Pearl, "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'01, Seattle, Washington: Morgan Kaufmann Publishers Inc., 2001, pp. 411–420, ISBN: 1558608001.

[78] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647. [Online]. Available: https://aclanthology.org/2020.acl-main.647.

[79] A. Kumar, C. Tan, and A. Sharma, "Probing classifiers are unreliable for concept removal and detection," in *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. [Online]. Available: https://openreview.net/forum?id=MozmMHehWW8.

[80] A. Amini, T. Pimentel, C. Meister, and R. Cotterell, *Naturalistic causal probing for morpho-syntax*, 2022. arXiv: 2205.07043 [cs.CL].

[81] A. Geiger, H. Lu, T. Icard, and C. Potts, "Causal abstractions of neural networks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 9574–9586. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

[82] A. Geiger, Z. Wu, H. Lu, *et al.*, "Inducing causal structure for interpretable neural networks," in *International Conference on Machine Learning*, PMLR, 2022, pp. 7324–7338.

[83] E. Goodwin, K. Sinha, and T. J. O'Donnell, "Probing linguistic systematicity," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 1958–1969. DOI: 10.18653/v1/2020.acl-main.177. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.177.

[84] A. Geiger, K. Richardson, and C. Potts, "Neural natural language inference models partially embed theories of lexical entailment and negation," *arXiv: Computation and Language*, 2020.

[85] N. F. Liu, R. Schwartz, and N. A. Smith, "Inoculation by fine-tuning: A method for analyzing challenge datasets," *CoRR*, vol. abs/1904.02668, 2019. arXiv: 1904.02668. [Online]. Available: http://arxiv.org/abs/1904.02668.

[86] J. Van Benthem *et al.*, *Essays in logical semantics*. Springer, 1986.

[87] H. Hu, Q. Chen, K. Richardson, A. Mukherjee, L. S. Moss, and S. Kuebler, "Monalog: A lightweight system for natural language inference based on monotonicity," in *Proceedings of the Society for Computation in Linguistics (SCiL) 2020*, 2020, pp. 319–329.

[88] L. Abzianidze, "A tableau prover for natural logic and language," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2492–2502. DOI: 10.18653/v1/D15-1296. [Online]. Available: https://www.aclweb.org/anthology/D15-1296.

[89] B. MacCartney and C. D. Manning, "An extended model of natural logic," in *Proceedings of the Eight International Conference on Computational Semantics*, Tilburg, The Netherlands: Association for Computational Linguistics, Jan. 2009, pp. 140–156. [Online]. Available: https://www.aclweb.org/anthology/W09-3714.

[90] G. Angeli and C. D. Manning, "Naturalli: Natural logic inference for common sense reasoning," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 534–545.

[91] D. Bobrow, B. Cheslow, C. Condoravdi, T. H. King, R. Nairn, and A. Zaenen, "Parc's bridge and question answering system," 2007.

[92] A.-L. Kalouli, R. Crouch, and V. de Paiva, "Hy-NLI: A hybrid system for natural language inference," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5235–5249. DOI: 10.18653/v1/2020.coling-main.459. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.459.

[93] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota:

Association for Computational Linguistics, Jun. 2019, pp. 4129–4138. DOI: 10.18653/v1/N19-1419. [Online]. Available: https://www.aclweb.org/anthology/N19-1419.

[94] L. Abzianidze, J. Bjerva, K. Evang, *et al.*, "The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 242–247. [Online]. Available: https://aclanthology.org/E17-2039.

[95] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[96] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[97] C. Herlihy and R. Rudinger, "MedNLI is not immune: Natural language inference artifacts in the clinical domain," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1020–1027. DOI: 10.18653/v1/2021.acl-short.129. [Online]. Available: https://aclanthology.org/2021.acl-short.129.

[98] Z. Chen, Q. Gao, and L. S. Moss, *Neurallog: Natural language inference with joint neural and logical reasoning*, 2021. arXiv: 2105.14167 [cs.CL].

[99] L. Abzianidze, "LangPro: Natural language theorem prover," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 115–120. DOI: 10.18653/v1/D17-2020. [Online]. Available: https://www.aclweb.org/anthology/D17-2020.

[100] K. Richardson, H. Hu, L. S. Moss, and A. Sabharwal, "Probing natural language inference models through semantic fragments," *CoRR*, vol. abs/1909.07521, 2019. arXiv: 1909.07521. [Online]. Available: http://arxiv.org/abs/1909.07521.

[101] E. Goodwin, K. Sinha, and T. J. O'Donnell, "Probing linguistic systematicity," *ACL 2020*, 2020.

[102] G. Alain and Y. Bengio, *Understanding intermediate layers using linear classifier probes*, 2018. arXiv: 1610.01644 [stat.ML].

[103] P. Lewis, P. Stenetorp, and S. Riedel, "Question and answer test-train overlap in open-domain question answering datasets," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 1000–1008. [Online]. Available: https://aclanthology.org/2021.eacl-main.86.

[104] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.703.

[105] D. Ferreira, J. Rozanova, M. Thayaparan, M. Valentino, and A. Freitas, "Does my representation capture X? probe-ably," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Aug. 2021, pp. 194–201. DOI: 10.18653/v1/2021.acl-demo.23. [Online]. Available: https://aclanthology.org/2021.acl-demo.23.

[106] K. Zhang and S. Bowman, "Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 359–361. DOI: 10.18653/v1/W18-5448. [Online]. Available: https://www.aclweb.org/anthology/W18-5448.

[107] J. Rozanova, D. Ferreira, M. Thayaparan, M. Valentino, and A. Freitas, *Supporting context monotonicity abstractions in neural nli models*, 2021. arXiv: 2105.08008 [cs.CL].

[108] J. Rozanova, D. Ferreira, M. Valentino, M. Thayaparan, and A. Freitas, "Decomposing natural logic inferences in neural NLI," *CoRR*, vol. abs/2112.08289, 2021. arXiv: 2112.08289. [Online]. Available: https://arxiv.org/abs/2112.08289.

[109] A. Ben-Israel, "Projectors on intersections of subspaces," *Contemporary Mathematics*, pp. 41–50, 2015.

[110] J. Rozanova, D. Ferreira, M. Thayaparan, M. Valentino, and A. Freitas, *Supporting context monotonicity abstractions in neural nli models*, 2021. arXiv: 2105.08008 [cs.CL].

[111] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv e-prints*, Feb. 2018. arXiv: 1802.03426 [stat.ML].

[112] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2126–2136. DOI: 10.18653/v1/P18-1198. [Online]. Available: https://aclanthology.org/P18-1198.

[113] A. Feder, N. Oved, U. Shalit, and R. Reichart, "CausaLM: Causal model explanation through counterfactual language models," *Computational Linguistics*, vol. 47, no. 2, pp. 333–386, Jun. 2021. DOI: 10.1162/coli_a_00404. [Online]. Available: https://aclanthology.org/2021.cl-2.13.

[114] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (cace)," *arXiv preprint arXiv:1907.07165*, 2019.

[115] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*, PMLR, 2018, pp. 2668–2677.

[116] I. Gat, G. Lorberbom, I. Schwartz, and T. Hazan, "Latent space explanation by intervention," *CoRR*, vol. abs/2112.04895, 2021. arXiv: 2112.04895. [Online]. Available: https://arxiv.org/abs/2112.04895.

[117] A. Abid, M. Yuksekgonul, and J. Zou, "Meaningfully debugging model mistakes using conceptual counterfactual explanations," in *International Conference on Machine Learning*, PMLR, 2022, pp. 66–88.

[118] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually augmented data," *International Conference on Learning Representations (ICLR)*, 2020.

[119] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial nli: A new benchmark for natural language understanding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020.

[120] Y. Nie, H. Chen, and M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[121] K. Richardson, H. Hu, L. S. Moss, and A. Sabharwal, "Probing natural language inference models through semantic fragments," in *AAAI Conference on Artificial Intelligence*, 2019.

[122] Y. Elazar, N. Kassner, S. Ravfogel, *et al.*, "Measuring causal effects of data statistics on language model's 'factual' predictions," 2022. eprint: arXiv:2207.14251.

[123] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? natural language attack on text classification and entailment," *arXiv preprint arXiv:1907.11932*, 2019.

[124] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4902–4912. DOI: 10.18653/v1/2020.acl-main.442. [Online]. Available: https://aclanthology.org/2020.acl-main.442.

[125] A. Feder, K. A. Keith, E. Manzoor, *et al.*, "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1138–1158, 2022. DOI: 10.1162/tacl_a_00511. [Online]. Available: https://aclanthology.org/2022.tacl-1.66.

[126] Z. Wu, K. D'Oosterlinck, A. Geiger, A. Zur, and C. Potts, "Causal proxy models for concept-based model explanations," 2022. arXiv: 2209.14279 [cs.LG].

[127] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, 2020.

[128] D. Hupkes, M. Giulianelli, V. Dankers, *et al.*, "State-of-the-art generalisation research in nlp: A taxonomy and review," English, ArXiv, WorkingPaper, Jan. 2023, We thank Adina Williams, Armand Joulin, Elia Bruni, Lucas Weber, Robert Kirk and Sebastian Riedel for providing us feedback on various stages of this draft, and Gary Marcus for providing detailed feedback on the final draft of this paper. We thank Elte Hupkes for making the app that allows searching through references, and we thank Daniel Haziza and Ece Takmaz for other contributions to the website. DOI: 10.48550/arXiv.2210.03050.

Appendix A

Code, Data and Models

Models

We use the following base models across the works in this thesis:

Base Model	Source
roberta-large-mnli	https://huggingface.co/roberta-large-mnli
bert-base-uncased-snli	https://huggingface.co/textattack/bert-base-uncased-snli
infobert	https://github.com/AI-secure/InfoBERT/tree/master/ANLI
roberta-large-snli_mnli_fever_anli	https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

The data and code for the associated fine-tuned versions is included in the repositories linked below.

Code and Data

We include the following repositories which contain the experimental code (including all hyperparameter configurations) and data used for the experiments in this thesis:

A.1 Supporting Context Monotonicity Abstraction in Neural NLI

The following repository includes our train/test split of the HELP dataset and the HELP-Contexts dataset:
https://github.com/juliarozanova/supporting_monotonicity

Decomposing Natural Logic Inferences in Neural NLI

The following repository incldes the NLI-*XY* dataset:
`https://github.com/juliarozanova/nli_xy`

Interventional Probing in High Dimensions: an NLI Case Study

`https://github.com/juliarozanova/mnestic_probing`

Causal Effects in Natural Logic Handling

`https://github.com/juliarozanova/nli_causal`