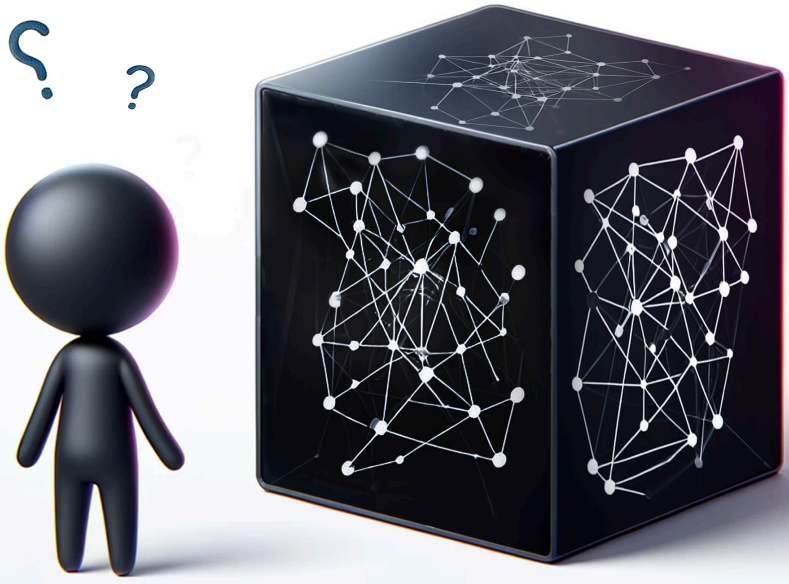
Linköping Studies in Science and Technology
Dissertation No. 2364

Understanding Large Language Models

Towards Rigorous and Targeted Interpretability Using Probing Classifiers and Self-Rationalisation

Jenny Kunz



LINKÖPING UNIVERSITY

Linköping Studies in Science and Technology
Dissertations, No. 2364

Understanding Large Language Models: Towards Rigorous and Targeted Interpretability Using Probing Classifiers and Self-Rationalisation

Jenny Kunz

Linköping University
Department of Computer and Information Science
Division of Artificial Intelligence and Integrated Computer Systems
SE-581 83 Linköping, Sweden

Linköping 2024

Typeset using LaTeX

Cover illustration by Max Trembczyk and Jenny Kunz, with assistance of the DALL·E 2 image generation model.

Printed by LiU-Tryck, Linköping 2024

Edition 1:1

ISBN 978–91–8075–470–5 (print)
ISBN 978–91–8075–471–2 (PDF)
https://doi.org/10.3384/9789180754712
ISSN 0345-7524

POPULÄRVETENSKAPLIG SAMMANFATTNING

Neuronala språkmodeller ligger bakom många vardagliga användningar, bland annat mjukvara som kontrollerar stavning och grammatik och som kompletterar text. De mest omtalade systemen just nu är dock chatbotar som ChatGPT. Sådana modeller har många imponerande färdigheter: Man kan ställa olika typer av frågor till dem, och svaren är ofta både korrekta och välskrivna. Chatbotar kan även hjälpa med textsammanfattningar, med att omstrukturera och förbättra text, med programmeringsfrågor och med många andra uppgifter.

En utmaning med dagens språkmodeller är att de är så stora och tränade på så pass stora mängder text att det är omöjligt att fullt ut förstå hur de fungerar och varför man får ett visst svar. Den största delen av sina färdigheter får modellerna från så kallad självövervakad träning: De är inte tränade på att svara på frågor utan att på att predicera nästa ord i en text (eller ord som saknas någonstans i texten) och lär sig själv vilken information som är viktig för att lösa den konceptuellt enkla uppgiften. Självövervakad träning har visat sig vara effektiv för att modellerna ska tillägna sig kunskap om lingvistik och fakta, men det är inte trivialt att veta exakt vilka kunskaper och egenskaper modellen får. En sådan förståelse är dock nödvändig för att kunna förbättra modellerna och för att bedöma när de är pålitliga och kan användas med gott samvete.

Syftet med den här avhandlingen är att utveckla, förbättra och utvärdera metoder inom förklarbar språkteknologi. Jag presenterar vårt arbete inom två delområden: interpretation av modellernas interna representationer och generering av förklaringar för individuella svar. I det första delområdet har vi utvecklat en metod för att förstå begränsningarna med en populär metod inom modellinterpretation som kallas för *probing*. Med hänsyn till dessa begränsningar har vi utvecklat metoder för att göra probing mer utmanande samt nya metriker för att mäta språkmodellers kvalitet. I det andra delområdet är fokusen på språkmodeller som har förmågan att kunna generera *förklaringar* för sina svar. Vi har systematiserat egenskaper som förklaringar som genereras av människor har och undersökt om de även finns i automatiskt genererade förklaringarna. Våra resultat visar att vissa egenskaper förekommer hos en stor del av de genererade förklaringarna, speciellt olika former av ofullständighet och illustrerande element. Subjektivitet förekommer mycket mer sällan i genererade förklaringar, antagligen för att ett senare steg i modellernas träningsprocess stävjar den här egenskapen. Mer generellt kan vi konstatera att mänskliga användare och tillämpningssystem som bygger på språkmodeller har olika behov med avseende på vilka egenskaper dessa förklaringar ska ha.

Vår forskning bidrar till en bättre förståelse av språkmodeller, men det är viktigt att vara tydlig med att vi är långt ifrån en detaljerad förståelse. Kompletterad med kunskap om modellens arkitektur, träningsdata och träningsmål kan kunskapen och metoderna dock vara till hjälp för att formulera och testa hypoteser för hur modellerna beter sig.

ABSTRACT

Large language models (LLMs) have become the base of many natural language processing (NLP) systems due to their performance and easy adaptability to various tasks. However, much about their inner workings is still unknown. LLMs have many millions or billions of parameters, and large parts of their training happen in a self-supervised fashion: They simply learn to predict the next word, or missing words, in a sequence. This is effective for picking up a wide range of linguistic, factual and relational information, but it implies that it is not trivial what exactly is learned, and how it is represented within the LLM.

In this thesis, I present our work on methods contributing to better understanding LLMs. The work can be grouped into two approaches. The first lies within the field of interpretability, which is concerned with understanding the internal workings of the LLMs. Specifically, we analyse and refine a tool called probing classifiers that inspects the intermediate representations of LLMs, focusing on what roles the various layers of the neural model play. This helps us to get a global understanding of how information is structured in the model. I present our work on assessing and improving the probing methodologies. We developed a framework to clarify the limitations of past methods, showing that all common controls are insufficient. Based on this, we proposed more restrictive probing setups by creating artificial distribution shifts. We developed new metrics for the evaluation of probing classifiers that move the focus from the overall information that the layer contains to differences in information content across the LLM.

The second approach is concerned with explainability, specifically with self-rationalising models that generate free-text explanations along with their predictions. This is an instance of local understandability: We obtain justifications for individual predictions. In this setup, however, the generation of the explanations is just as opaque as the generation of the predictions. Therefore, our work in this field focuses on better understanding the properties of the generated explanations. We evaluate the downstream performance of a classifier with explanations generated by different model pipelines and compare it to human ratings of the explanations. Our results indicate that the properties that increase the downstream performance differ from those that humans appreciate when evaluating an explanation. Finally, we annotate explanations generated by an LLM for properties that human explanations typically have and discuss the effects those properties have on different user groups.

While a detailed understanding of the inner workings of LLMs is still unfeasible, I argue that the techniques and analyses presented in this work can help to better understand LLMs, the linguistic knowledge they encode and their decision-making process. Together with knowledge about the models' architecture, training data and training objective, such techniques can help us develop a robust high-level understanding of LLMs that can guide decisions on their deployment and potential improvements.

Acknowledgments

Since I came to Linköping five years ago, a lot of people have contributed to the work on this thesis and to making my PhD experience enjoyable and fun.

First and foremost, I want to thank my supervisor Marco Kuhlmann for your support and guidance. It's been super inspirational to work with someone with such a commitment to rigorous research practice, attention to detail, and not least a passion for great teaching.

A huge thank you to my other colleagues in the NLP group, Ehsan Doostmohammadi, Oskar Holmström, Olle Torstensson, Marcel Bollmann, Kevin Glocker and Noah-Manuel Michael, for sharing ideas and experiences and for discussions about research and everything else. I also want to thank my amazing master's thesis students Martin Jirénius and Marc Braun for their great work on their thesis projects and on the papers we have co-written. And not to forget all my former colleagues from NLPLAB, Arne Jönsson, Lars Ahrensberg, Robin Kurtz, Evelina Rennes, Jody Foo, Jalal Maleki and Riley Capshaw, and my co-supervisor Eva Blomqvist. You all provided invaluable feedback and inspiration especially at the beginning of my PhD journey.

I thank Richard Johansson, Lovisa Hagström and Tobias Norlund for many research discussions especially during the pandemic and when our own NLP group was still tiny. Thanks to Ryan Cotterell for the insightful discussions in my mid-term seminar, and to everyone else who attended my seminars and presentations over the years and gave feedback to my work.

All the lunch and fika breaks wouldn't have been as fun without the guys (m/f/d) from the planning group and elsewhere at AIICS and HCS. Thank you for your company!

I thank Karin Baardsen for your assistance with everything concerning travel and money, and Anne Moe for your great support with the bureaucratic processes and everything else. Thanks to my colleagues in the PhD Council for constantly working on improving the PhD experience at IDA.

A great thank you to my family for supporting me and my dreams and ideas, even though they tend to send me far away from you. And finally, a huge thank you to my three favourite people: To my partner Max for coming to Linköping with me and for all the support over the years, and to our kids Josefine and Rufus for making life colourful. I love you!

Contents

List of Figures

List of Tables

Part I

Introductory Summary

1 Introduction

This chapter includes the motivation behind the work in this thesis (Section 1.1), its scientific contributions (Section 1.2) and its delimitations (Section 1.3), as well as an outline of the thesis structure (Section 1.4).

1.1 Motivation

Over the course of the recent five to seven years, large language models (LLMs) have become almost inevitable in natural language processing (NLP). LLMs are huge neural networks with many millions or billions of parameters that learn to represent language by predicting words from their context. By being trained on large amounts of texts and making such predictions of words billions of times, LLMs learn to capture fine-grained statistical information about language. These features have proven proven useful for almost all kinds of downstream applications that are currently of interest in NLP. LLMs are even applied to syntactic tasks but stand out in particular for natural language understanding (NLU) tasks that depend on complex interactions between words and sentences, as well as on commonsense facts about the world we live in. Many NLU benchmarks that were once considered challenging, such as extractive question answering (Rajpurkar et al. 2016)[1] and the recognition of textual entailment (A. Wang et al. 2019)[2], can now be seen as solved: Current models regularly surpass human performance, although it is important to emphasise that the models perform well

[1] `https://rajpurkar.github.io/SQuAD-explorer/`; last accessed 21/2/2024.
[2] `https://super.gluebenchmark.com/leaderboard`; last accessed 21/2/2024.

on *benchmarks* rather than real-world *tasks*, and that the validity of human-to-system comparisons has been challenged (Tedeschi et al. 2023).

The popularity of LLMs comes from performance gains in many tasks, but also from the fact that they make careful feature engineering obsolete as they already appear to contain a lot of useful syntactic, but also semantic information. This broad applicability has brought the first LLMs such as *ELMo* (Peters et al. 2018) and *BERT* (Devlin et al. 2019) wider media attention. In 2018, *The New York Times* wrote in an article about BERT that "*computer systems can learn the vagaries of language in general ways and then apply what they have learned to a variety of specific tasks.*"[3]. The article even noted BERT's human-like performance on a commonsense reasoning task.

More recently, the focus has shifted towards *generative* LLMs such as *GPT-4* (OpenAI 2023) and *Llama 2* (Touvron et al. 2023). These models do not only solve classification and labelling tasks but can generate coherent text for a wide range of user prompts. The largest of these models can do so even without adaption to the specific task or with only few or no task-specific examples (Brown et al. 2020; Chowdhery et al. 2022). With the arrival of easy-to-use conversational interfaces such as OpenAI's *ChatGPT*, Google's *Bard* and Anthropic's *Claude*, generative LLMs have been widely adopted even by non-NLP people. In a survey of university students in Sweden, Malmström et al. (2023) report that 63% of the respondents report to use ChatGPT at least occasionally, and only 5% are completely unfamiliar with it. While sampling bias may apply, and university students may be more likely to adopt new technologies than the general population, this nonetheless indicates that LLMs are impacting society more and more, with no end in sight.

As LLMs are adopted by broader parts of the population and for an increasing number of use cases, it is crucial to understand how they work and what they have learned. However, much of the nature of the features that LLMs encode remains unclear even to NLP experts. Neural network-based models in general, and LLMs in particular, are opaque. Their size and complexity make a complete, fine-grained understanding of the internal processes infeasible. For this reason, they have unexpected failure modes (Bommasani et al. 2021; Mittelstadt et al. 2019). This affects the users' trust in a system and the ability of operators to know when it is a good idea to give a system control (Lipton 2018), but also the developer's possibilities to improve it as a limited understanding of models also hinders hypothesis-driven progression (Rogers et al. 2020). Without knowledge about what features a model relies on, it is also hard to assess how fair and ethical its decisions are (Miller 2019). It also leaves open the question of the true capabilities of current NLP models: The great performance on many tasks could be (and in some cases has been) attributed to statistical cues and biases in the data sets instead of models performing the reasoning that we hope it to do. Besides developing more challenging tasks and data sets, we also need more insights about the LLMs' inner workings in order to get a more robust understanding.

For these reasons, an increasing effort within the NLP community is spent on the interpretability and explainability of LLMs. Interpretability methods help to develop a

[3] `https://www.nytimes.com/2018/11/18/technology/artificial-intelligence-language.html`; last accessed 15/2/2024.

higher-level understanding of the LLMs' internal representations and test hypotheses about their inner workings. Explainability methods showcase (actual or potential) reasons that led to specific predictions.

While many ideas, approaches and framings of interpretation techniques pop up, their methodology remains largely explorative. Oftentimes, the approaches are limited to very focused areas, or build up on cherry-picked examples. A more systematic, rigorous and interpretable approach to interpretability is needed for faithful, reliable insights. In this thesis, I present our work on these issues: on better understanding the contributions of current interpretability and explainability methods, and on making them more rigorous. I argue that such approaches, paired with an understanding of the models' architecture, learning objective(s) and training data, can lead us to a level of understanding that allows us to predict the models' behavior and outputs of future tasks. In short, the goal is to *demystify* LLMs.

1.2 Contributions

This thesis is divided into two parts, which represent two complementary approaches to better understanding LLMs and their decision-making process: interpretation and explanation. As those approaches contribute to different subgoals of explainable NLP and use different methods, I will first introduce and discuss them separately, before bringing them together again in Chapter 6, where I will discuss how the combination of both approaches helps us to understand the bigger picture of how LLMs work.

1.2.1 Interpretation

In the first part of this thesis, I investigate how linguistic information is structured within the model by focusing on the roles that the internal representations at the various layers of the models play. This gives us an improved understanding of the model on a *global* level: We get to understand how the model as a whole represents language. We aim to answer the following research questions:

1. What limitations do currently popular interpretation techniques and explanation mechanisms have? How can we assess and expose the weaknesses of such mechanisms?
2. How can interpretation methods be improved so that they are more faithful, more reliable and better suited to draw conclusions about what the language representations encode and how information is structured within the models?

To answer these questions, our work assesses, improves, and develops the interpretation methodologies. It focuses on a popular interpretability tool that is called a *probing classifier*. A probing classifier is a tool that learns to predict linguistic properties from the internal representations of LLMs and other neural network-based models. An important but non-trivial distinction is if the classifier *learns* the probing task from the data it is trained on, either based on memorisation or on contextual patterns, or if it *extracts* relevant linguistic information that is encoded in the LLM's representations.

Popular interpretation methodologies are questioned and theoretically and empirically tested on whether they really reveal what they intend. In Paper I, *Classifier Probes May Just Learn from Linear Context Features*, we develop a framework that exposes that an important assumption made in previous work does not hold neither logically nor empirically, namely, that if we feed the representation of one word at a time to the probing classifier, it does not have sufficient contextual information to learn the task. We show that information about each word's surrounding context is extractable from its own representation with impressive precision, and that none of the baselines or restrictions commonly used in the literature can disprove the hypothesis that it is only linear context information that enables the probing classifier to learn the task. In conclusion, common probing methodologies are *unable* to differentiate between *learning* the task and *extracting* relevant features.

Based on our findings from Paper I, we increase the restrictiveness of the probing setup in Paper II *Test Harder than You Train: Probing with Extrapolation Splits*. Instead of evaluating the probe on a similar distribution to the one it was trained on, we create extrapolation splits: We define a set of criteria to rank the difficulty of data samples, and use only the simpler samples for training and the harder samples for evaluation. As machine learning models, of which the probing classifier is an instance, typically only succeed in an interpolation setup, we argue that success in the extrapolation setup is a sign that the probing classifier *decodes* task-specific linguistic properties from the representations rather than learning the task.

In Paper III, *Where Does Linguistic Information Emerge in Neural Language Models? Measuring Gains and Contributions across Layers*, we overcome the problems that we identified in Paper I with new metrics that provide a new perspective on the structure of information within models. Instead of comparing the probing classifier's results on a specific layer to a global baseline, we focus on local gains: That is, how much information the representation at this layer contains that the representation of the preceding layer did not contain. We show that this perspective changes the focus for syntactic tasks from the middle layers, where the overall performance is highest, to the first layers, where most new information emerges.

1.2.2 Explanation

In the second part of this thesis, I will turn to models that explain their own individual predictions, so-called self-rationalising models. I focus on models that generate those predictions in free text as this form is both easily understandable to various user groups and applicable to many kinds of NLP tasks. As those generated explanations are however not (necessarily) faithful to the predictions but generated by an opaque model themselves, our work aims to better understand their properties. In particular, it focuses on the relation between various, partially conflicting goals of explainable NLP: That the explanations should provide us with insights about the LLM's true prediction process, guide the LLM itself in its predictions, but also be understandable and useful to different users, such as non-technical end users, but also developers.

In this part, our work investigates the following research questions:

1. How does the utility of explanations to a downstream model align with the human perception of the explanations?
2. Which properties of human explanations does an LLM adopt and how do these relate to different goals of explainable NLP?

These questions are investigated in two papers that are centred around human evaluations of LLM-generated explanations.

In Paper IV, *Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions*, we answer the first question by comparatively training models with different pipeline architectures and performing human evaluations on them. We show that a crucial difference is that a downstream model by tendency benefits from the inclusion of novel (input-external) information in the explanation, even if the information in many cases is factually incorrect, while human raters however punish factual incorrectness decisively.

In Paper V, *Properties and Challenges of LLM-Generated Explanations*, we perform a human evaluation on an LLM's outputs for a diverse dataset. We annotate the data for known properties of human explanations, specifically such that have been pointed out as disadvantageous for explainable NLP. We find that the LLM explanations are often incomplete and contain illustrative elements, but are rarely subjective and rarely misleading, but that the latter two findings are highly dependent on the model and dataset that we use. We connect the observed properties with different goals and user groups of explainable NLP, showing that all of the properties can have positive or negative implications depending on the use case.

1.3 Delimitations

The aim of this thesis is not to develop fully interpretable language models where the internal decision steps can be understood in detail. Classical explainable artificial intelligence methods such as sparse representations, few learnable parameters, or decomposability are not considered. I accept the dense and generally opaque nature of the currently most successful representations and explore and develop methods to understand their properties better. The understanding I aim at is at a relatively high level. While the methods used may help to predict the success and failure cases of models, and add context for human users when making decisions, they do not make the model a reliable basis for high-stake decisions without a human expert in the loop.

1.4 Reading Guide

This thesis assumes its readers to have a basic understanding of machine learning and natural language processing concepts. It does not require a deep technical understanding of current models; I introduce them at the level of details that is needed to follow our work.

Outline. This thesis is written in form of a *compilation*. The original research conducted as part of this thesis is described in the five articles found in **Part II**, in their

published (Paper I–IV) or submitted (Paper V) form. **Part I**, the introductory summary (*kappa*), provides a more comprehensive background, a broader overview of relevant literature and a more extensive discussion of the methodologies that are the basis of our own work. The kappa is structured as follows:

- In **Chapter 2**, I give readers an overview of the field of language representation learning. In particular, I introduce the three types of representations that our papers build on: the recurrent neural network-based ELMo, the Transformer encoder-based BERT and the family of Transformer decoder-based GPT models. As those models are widely known in the NLP community, readers with this background may skip this chapter.
- In **Chapter 3**, I introduce interpretation techniques that are designed to assess what is learned by the LLMs. I briefly introduce a broader set of methods to position our contributions in the field as a whole but focus on probing classifiers that are the method used in our own work.
- In **Chapter 4**, I introduce work on generating and evaluating explanations in NLP. I start with an overview of the types of explanations that exist in NLP but focus on free-text explanations that are used our work.
- Our contributions to the fields of interpretability and explainability are then summarized in **Chapter 5**, where I give an overview of the findings of the papers that this thesis contains and relate them to each other. I also introduce further papers I contributed to during my time as a PhD student.
- **Chapter 6** concludes this thesis with a final summary and discussion of our contributions as well as my view on the future of the field.

Throughout the thesis, I use the pronoun *I* when referring to the writing of this kappa. Whenever referring to work done with collaborators, I use *we*.

2 Language Representation Learning

The object of study of this thesis are LLMs, the currently most dominant type of neural language representations in academic NLP. In this chapter, I give an overview of the fast progress in language representation learning over the past ten years.

Language representation models automatically identify and organize re-usable information from text corpora to build a representation of natural language. Those representations can be built either for internal use in the model itself based on the task-specific dataset that this model is trained on (Section 2.1) or from unlabelled data for the transfer to various models (Section 2.2) or tasks (Section 2.3). The most weight will lie on the latter representations, as the LLMs that we study in our experiments are instances of those. The development of interpretability techniques was however already taking off in the context of pre-LLM neural models and word embeddings, as we will discuss in Chapter 3. After having introduced the models, I will conclude the chapter by discussing some crucial limitations of current LLMs in Section 2.4.

2.1 Neural Network Language Representations

In NLP, neural networks gained broad popularity in the middle of the last decade. It was recurrent neural networks (RNNs; Elman 1990), and long short-term memory networks (LSTMs; Hochreiter and Schmidhuber 1997) in particular, that first led to major performance improvement on many tasks. The sequential nature of these architectures was a natural fit for language data, which is also generated and processed sequentially by humans. In addition to performance gains, neural networks dispensed with the need for manual feature engineering and hand-crafted pipelines by building

up their own internal language representation. In the pre-neural era, even shallow tasks like syntactic dependency parsing relied on a large set of features, such as part-of-speech (POS) tags of the current tokens themselves and surrounding tokens, and various distance- and direction-based features. Finding the best set of features was an important and time-consuming part of parser engineering. Kiperwasser and Goldberg (2016) in their influential parser use a bidirectional LSTM encoding to represent a token in its context, and base the representation only on word forms and POS tags, and let the LSTM encoder do the rest. But even POS tags were no longer a crucial advantage for syntactic parsing, giving at most a slight performance gain (Dozat et al. 2017; Lhoneux et al. 2017). Given enough training data, the models can operate on raw text without helper systems, and achieve similar accuracy to highly engineered models. A similar transformation happened in many other tasks, with an important source of influential new ideas being machine translation that brought along RNN-based sequence-to-sequence approaches (Sutskever et al. 2014), attention (Bahdanau et al. 2015), and finally the Transformer architecture (Vaswani et al. 2017) that is the currently most widely adapted base model in NLP. Learning to encode a word with the task-specific data set, and to contextualize it, with a neural model proved to be a very successful representation for language. But while this is a form of feature learning to get a model-internal language representation, such representations are still trained task-specifically (although the word representation architecture can be shared among many tasks), and usually are not transferred to other purposes.

2.2 Word Embeddings

At the same time as neural networks were popularized in NLP, word embedding (WE) models appeared as widely used general-purpose word representations. They build one static vector for each word form based on its immediate contexts occurring in an unannotated corpus, using a simple word prediction task based on the context words or on a dimensionality-reduced co-occurrence matrix. Popular examples are Turian embeddings (Turian et al. 2010), GloVe (Pennington et al. 2014) and the word2vec models (Mikolov et al. 2013b). Word embeddings can be applied to basically every NLP task and in any machine learning model. Using them improved the neural network performance on many tasks due to their richer, self-supervised representation of words, in particular on tasks with limited data that is not sufficient for the model to build its own rich language representation.

2.3 Contextualized Language Representations

While the word embeddings described in Section 2.2 were used as the input to other neural network-based models, large contextualized language representations transformed the field and how models are built fundamentally. Pre-trained on huge amounts of unlabelled text data and with rapidly increasing model sizes, such a model can with a relatively modest effort be adapted to various kinds of unforeseen tasks. The step of adapting such a model with task-specific data in a second, substantially shorter training phase is called *fine-tuning*. For many tasks, they easily outperform previous neural network models trained from scratch. As Bommasani et al. (2021) note, "*the field of NLP has become largely centred around using and understanding foundation*

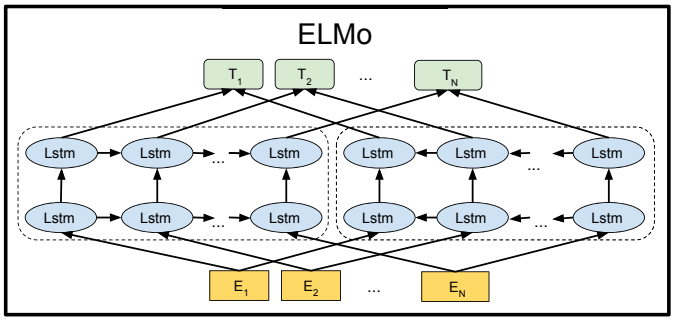


Figure 2.1: Overview of ELMos architecture. Figure by Devlin et al. (2019).

models". Few pre-trained models are re-used and examined tens of thousands of times, with the BERT model currently (as of February 2024) having almost 100,000 citations on Google Scholar.

Contextualized language representations are largely congruent with LLMs. While the latter term is more often used for generative models, I adopt the definition of Luccioni and Rogers (2023) which states that LLMs are neural network models that process and generate text, that have been trained on at least one billion words, and that make inference based on transfer learning. While there are contextualized language representation that do not fulfill the data requirements of this definition, we see in Table 2.1 that all of the language representations that were used in the experiments for this thesis project fulfill it. I will therefore use the term LLMs interchangeably.

Model	Architecture	Objective	#Params	Training Data	Papers
ELMo	BiLSTM	Autoreg. LM	94M	1B Words	I
$\text{BERT}_{\text{base}}$	Tf. Encoder	MLM / NSP	110M	3.3B Words	I–IV
$\text{BERT}_{\text{large}}$	Tf. Encoder	MLM / NSP	340M	3.3B Words	-
GPT-2	Tf. Decoder	Autoreg. LM	1.5B	9B Tokens	IV
GPT-3	Tf. Decoder	Autoreg. LM	175B	300B Tokens	-
GPT-4	?	?	?	?	V

Table 2.1: Architecture, number of parameters and training data size for the LLMs used in our experimental work. For GPT-4, there is no official data available.

2.3.1 ELMo

The autoregressive language model ELMo (Peters et al. 2018) was, along with ULMFiT (Howard and Ruder 2018), one of the first contextualized word representation that was widely adopted in the NLP community. ELMo is intended and mostly used as a word representation in a downstream model. In contrast to the representations introduced in Section 2.2 however, the representation is a function of the entire input sentence rather than of a word.

ELMO is based on a bidirectional LSTM (Hochreiter and Schmidhuber 1997) consisting of a forward and a backward language model that predicts the next (respectively the previous) token conditioned on the LSTM accumulation of the preceding (respectively the future) tokens, maximizing the log-likelihood of both directions. For a sequence of N tokens $(t_1, t_2, \ldots, t_N)$ and the parameters Θ_x of the token representation, $\overrightarrow{\Theta}_{LSTM}$ of the forward-pass LSTM predicting from the left-side context, $\overleftarrow{\Theta}_{LSTM}$ of the backward-pass LSTM predicting from the right-side context, and Θ_s of the softmax layer, ELMo maximises:

$$\sum_{k=1}^{N} (\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s)) .$$

ELMO has a character-based word representation layer with 512 dimensions and 2 bi-LSTM hidden layers with 1,024 units. The token representation commonly used in downstream tasks is either the top layer or a task-specific weighted sum of the 3 internal layers of the LSTM. In the latter case, the weights for each layer are learned by the downstream model, while the parameters of the layers themselves remain frozen.

ELMo is trained on the One Billion Word Benchmark, a sentence-level English-language dataset in the news domain with, as the name says, approximately one billion words (Chelba et al. 2013).

2.3.2 BERT

BERT (Devlin et al. 2019) is based on a Transformer model's encoder (Vaswani et al. 2017) that contextualizes the word representations with multi-head self-attention and fully connected layers. Its architecture became the basis of many more Transformer-based language representations. Transformers proved to be more successful than other architectures like RNNs because they scale up to very deep models and the self-attention makes them more successful at catching long-range interactions of tokens (Bommasani et al. 2021).

After the input to the BERT model is tokenised, the embedding of the token itself is enriched with an encoding of its position in the input span, as the Transformer architecture does not natively model word order, and an encoding indicating which sentence the token belongs to, as BERT can process up to two sentences at a time. The standard $\text{BERT}_{\text{base}}$ model consists of 12 layers, while the $\text{BERT}_{\text{large}}$ model has 24 layers. The core components of each layer are a multi-head self-attention module and a fully connected layer. The self-attention module provides the representation with context, as it aggregates information from the embeddings of the whole input sequence. This is done multiple times in parallel (12 times in the case of $\text{BERT}_{\text{base}}$; 16 times for $\text{BERT}_{\text{large}}$) to be able to capture richer features from the representations (thus *multi-head* self-attention). The scores of each head are then combined before they are propagated to the fully connected layer. A high-level illustration of the architecture can be found in Figure 2.3.

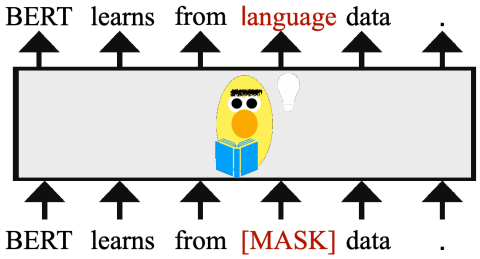


Figure 2.2: Example for MLM prediction: The word *language* is predicted by the BERT model based on its surrounding words.

BERT is pre-trained with two objectives: The Masked Language Model (MLM) randomly replaces some input tokens with a special MASK token, with the objective of predicting the vocabulary ID of the original token at that position. Figure 2.2 shows an example of such a prediction. This approach naturally includes both left and right context. The Next Sentence Prediction (NSP) objective makes BERT learn the relationship between two sentences by predicting if the second sentence is following the first one in the original document or not. A special token (*[CLS]*) is added for the NSP objective, which can also be used for other downstream classification tasks.

The MLM objective and the self-attention mechanism allow BERT to *jointly* condition the representation on the right and left context of a token, unlike ELMo that models them with two separate LSTMs. BERT thus optimises a simpler expression than ELMo. For N tokens $(t_1, t_2, \ldots, t_N)$ and the learnable parameters Θ, we get:

$$\sum_{k=1}^{N} \log p(t_k \mid t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_N; \Theta)$$

BERT is trained on two corpora: the BooksCorpus (Zhu et al. 2015), consisting of books with 800 million words, and English Wikipedia, consisting of 2.5 billion words.

2.3.3 GPT-2

The GPT model family consists of autoregressive LLMs based on the *decoder* part of the Transformer architecture. In contrast to the Transformer encoder that is the architecture of BERT, the decoder is designed for *text generation*. Therefore, it computes constraint multi-head self-attention scores that are only based on the context on the left of the token that is to be predicted. The effect of this constraint, as compared to BERT's Transformer encoder architecture, can be seen in Figure 2.3.

The conditioning on the left context only gives us the following formula (again, for a sequence if N tokens $(t_1, t_2, \ldots, t_N)$ and the parameters Θ):

$$\sum_{k=1}^{N} \log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta)$$

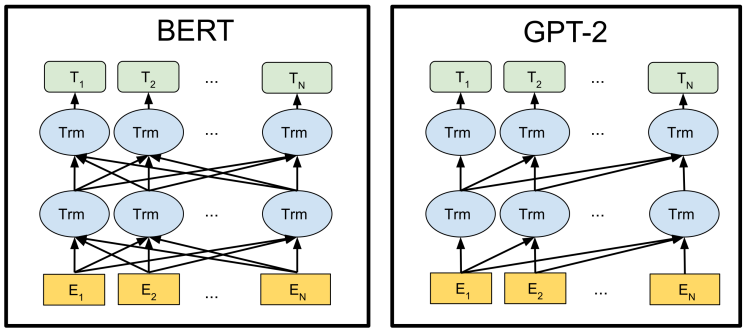


Figure 2.3: Comparison of the high-level architectures of BERT and GPT-3: The constrained multi-head self-attention of the Transformer decoder caps the connections to the preceding tokens in GPT-2. Figure by Devlin et al. (2019).

GPT-2's architecture and objective were introduced for the original GPT model by Radford et al. (2018). Its successor, GPT-2 (Radford et al. 2019), was the last GPT model with a full public release of the parameters. There are GPT-2 models in four different sizes, with the smallest one, like BERT, consisting of 12 layers and 12 self-attention heads, and the largest of 48 layers and 25 attention heads.

GPT-2 is trained on the WebText corpus that is also introduced in Radford et al. (2019). WebText is a scrape of outbound links from the internet forum Reddit. Radford et al. (2019) state that it contains approximately 8 million documents, and 40 gigabytes of text. While the original dataset is not shared and the exact number of training tokens therefore unknown, there exists an open replication by Gokaslan et al. (2019) which has 9 billion tokens.

2.3.4 GPT-3

GPT-3 (Brown et al. 2020) adopts GPT-2's architecture and objective, but it represents an enormous upscaling, growing from (at most) 1.5 billion to 175 billion parameters for the largest of the 8 GPT-3 models. The 175 billion parameters are spread over 96 layers, and the model has 128 attention heads.

The most groundbreaking property of GPT-3 is the ability to do *few-shot learning*, also called *in-context learning*: To adapt the model to a new task, no parameter updates are necessary; providing a limited number of task demonstration examples in the input is sufficient. Some success is even reported for *one-shot* and *zero-shot* transfer where only one or no demonstration example is provided to the model when solving a new task. It appears that, with sufficient scale, autoregressive pre-training is sufficient to infer the structure of many tasks. It also improved the performance over smaller models. However, it is well-documented that the training corpus of GPT-3 has a significant amount of contamination with common benchmark tasks (Brown et al. 2020; Dodge et al. 2021), which has a measurable effect on the performance of models (Magar and Schwartz 2022). Therefore, comparisons have to be done with caution.

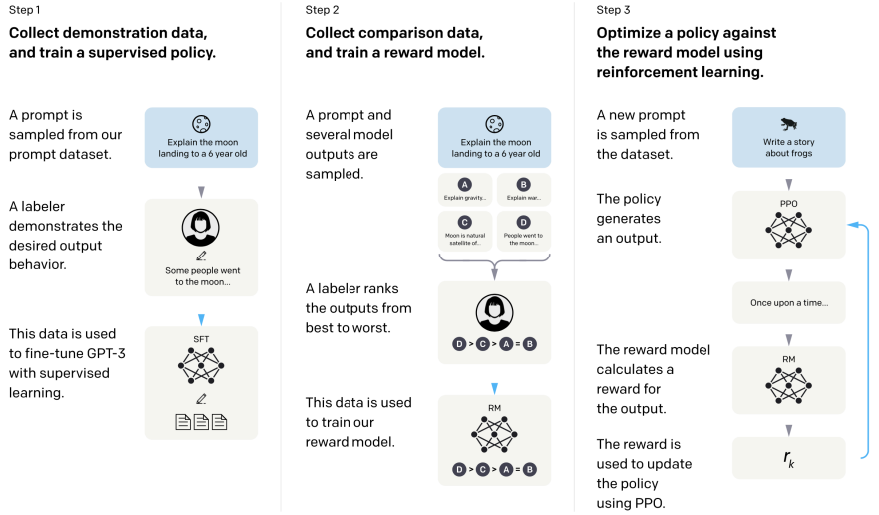


Figure 2.4: The three steps of aligning an LLM: Instruction fine-tuning, training the reward model, and reinforcement learning. Figure taken from Ouyang et al. (2022).

The training corpus for GPT-3 has 300B tokens and consists of is a filtered version of the web archive corpus CommonCrawl (Raffel et al. 2020), WebText2 (an expanded version of the GPT-2 training set), two internet-based corpora named Books1 and Books2 (with no further details released), and English-language Wikipedia.

2.3.5 GPT-3.5 and GPT-4

GPT-3.5 and GPT-4 (OpenAI 2023) are, as the names imply, the successors of GPT-3, and the base of the now-famous *ChatGPT* model. They have greatly improved zero-shot capabilities compared to GPT-3, probably due to *instruction fine-tuning* (IFT; Wei et al. 2021) and *reinforcement learning from human feedback* (RLHF; Ouyang et al. 2022). Those techniques can teach a model to better follow instructions and align them with the users' expectations. After the general language modelling pre-training phase, the model is fine-tuned on a supervised dataset containing a diverse set of instructions and demonstrations of desired model behaviours. With a sufficiently diverse set of tasks in the instructions, the model's performance is increased even on tasks unseen during IFT (Wei et al. 2021). In the RLHF step, human ratings of model outputs are used as a reward signal to the model. A reward model is trained on human rankings of multiple output candidates, which is then used to train the model via reinforcement learning. An overview of the process as introduced in the OpenAI's InstructGPT paper (Ouyang et al. 2022) is given in Figure 2.4. The whole process is also called *alignment*, as it does not only improve the zero-shot performance of the model but also aligns its outputs better with human expectations.

However, which techniques are used for GPT-3.5 and GPT-4 is speculative. While key information about the GPT-3 model, like the type and amount of the training data

and the model architecture, was still published, for the following models there is no reliable data available.

2.4 Limitations of Large Language Models

While LLMs have impressive coverage and adaptability, they continue to have fundamental limitations. When the task at hand is specialised and a sufficient amount of training data is available, it is often preferable to use a smaller, specialised model rather than a very large general-purpose model. One obvious reason is that the inference costs are substantially lower, but fine-tuning a smaller model may also result in better performance. The paper *Jack of all Trades, Master of None* by Kocoń et al. (2023) shows that the GPT-3.5-based ChatGPT model is often outperformed by task-specific fine-tuned models.

While this factor may change at any moment, there are more systematic limitations inherent to LLMs that I will summarise in this section: The generalistion capabilities, hallucinations, and the aspects of meaning they can capture. All of those limitations also relate to the lacking interpretability of the models. They show us that our understanding about (and control of) how LLMs model language and knowledge, and how they solve specific tasks, is still limited.

2.4.1 Generalisation

The ability to generalise, that is, to transfer representations, knowledge and strategies to new tasks, is a key goal of machine learning (Hupkes et al. 2022). Yet, LLMs have deficiencies in various types of generalisation.

They are prone to overly relying on superficial cues and annotation artifacts from the dataset when making predictions. In the famous paper *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*, T. McCoy et al. (2019) introduce an evaluation set for the sentence-level entailment prediction task NLI with examples where simple syntactic heuristics fail. While BERT on average performs better than the other three models that are trained from scratch, the performance is still poor compared to the standard evaluation set, indicating that even BERT relies too much on shallow heuristics rather than learning proper generalizations. The generalization problem is also addressed by Niven and Kao (2019). They apply BERT to an argument reasoning comprehension task and reach 77%, which is almost human performance and should the authors' view not be possible without supplying world knowledge. However, thet show that BERT relies largely on very simple statistical cues like unigrams and bigrams. As in the previously mentioned work however, this applies even stronger to models trained from scratch. They create an adversarial dataset by negating the claim and inverting the label of each data point, with the result that BERT performs only slightly above the random baseline. E. Wallace et al. (2019a) create adversarial examples with simple modifications of the inputs that change the prediction for several tasks and models including ELMo and GPT-2. The adversarial modifications let the performance drop substantially. Similarly, Hsieh et al. (2019) employ five different strategies for developing adversarial examples that could mislead neural models but not humans and find out that Transformer and BERT

models are less sensitive to them than recurrent models. Building on such works, Ilyas et al. (2019) argue that adversarial vulnerability is a natural consequence of the supervised paradigm. Models are trained on standard datasets containing signals that are incomprehensible to humans but highly predictive for the given dataset, which they call *non-robust features*. They emphasise the need to include human priors at training time to align the models with human expectations, and to make them robust and interpretable.

Larger LLMs have become less reliant on specific properties of the dataset as the need for fine-tuning with extensive task-specific datasets that may contain annotation artifacts has decreased. However, they are still heavily reliant on the presence of similar text in the pre-training data. R. T. McCoy et al. (2023) show that for exactly the same task, high-probability input or output sequences lead to much higher performance than equivalent low-probability sequences. For example, for linear functions, functions that occur often in text data because they e.g. are used for Celsius-to-Fahrenheit conversion lead to much more accurate results than a very similar function that uses only slightly different numbers in the same tasks. Sun et al. (2023) show that instruction-tuned LLMs are sensitive to instruction phrasing: They considerably drop in performance even with slight variations of the instructions, indicating that even the most advanced LLMs still struggle heavily with generalisation.

2.4.2 Hallucinations

The term *hallucinations* refers to information that is made up by the LLM. Coined in the field of neural machine translation (K. Lee et al. 2018) and subsequently adapted to abstractive text summarisation (Maynez et al. 2020), the term was originally used for information that is not faithful to the input document: The translation or the summary contain information that has not been present in the original text. Today however, it is frequently applied broadly in the field of text generation, referring to the inclusion of unintended information in general, such as factually incorrect or irrelevant information (Ji et al. 2023). As LLMs often operate in contexts where there is no single input document that the generated text should be closely aligned with, it is harder to define or detect a hallucination. It is however a common observation when working with LLMs that the output is plausible, but does not align with the real world. A hallucinated, i.e. factually incorrect, answer can even come with a *hallucinatory explanation*, a potentially plausible-sounding defense of such a statement (Augenstein et al. 2023).

Hallucinations are a major problem of LLMs not only because they decrease the performance of a system but also because they can pose a safety risk in sensitive applications, which limits the potential applications LLMs can realistically have (Ji et al. 2023). Therefore, reducing them has become the motivation behind a large body of research. Various techniques have been proposed for this purpose, such as confidence-based refinement of the output (Nie et al. 2019), refinement by checking against facts in a knowledge graph (Dziri et al. 2021) or retrieval-augmented generation (Shuster et al. 2021). However, hallucinations will not be fully mitigated in LLMs as we currently understand them. Mechanisms that rely on external ground truths such as knowledge graphs or text retrieval databases will not have full coverage over

all types of information that users of a general-purpose LLM such as GPT-4 seek. Reference-free approaches, such as the confidence-based refinement, can never capture all cases of hallucination, not least because making up content is be an inherent (and often intended) functionality of generative LLMs.

2.4.3 Form and Meaning

A more philosophical debate addresses the question if LLMs are even capable of capturing meaning. The LLMs we use build on the distributional hypothesis after which words that occur in similar contexts have similar meanings (Firth 1957; Harris 1954). The distributional approach has limitations in which aspects of meaning it can cover, in particular, it does not have access to the referential meaning of a word (Emerson 2020), meaning that they cannot relate the form of language they get to see to any instances of the real world. Bender and Koller (2020) argue that LLMs are therefore unable to achieve a semantic understanding of language.

The view that reference determines meaning is however disputed, especially because humans regularly use terms that have no referent in the real world because they refer to abstract concepts or invented entities. Piantadosi and Hill (2022) argue that LLMs capture other key aspects of meanings: *conceptual role meanings*, i.e. meanings derived from the relations of concepts to each other. They argue that reference is not necessary to determine meaning as humans can readily reason about many concepts without referents in the real world. Even Mollo and Millière (2023) emphasise that human representations often emerge without contact to a reference but by testimony based on the representation of others. They criticise the conflation of *grounding* and other concepts like *understanding* or *agency*, and define grounding as the ability to represent things independently of human interpretation, i.e. that they possess *intrinsic meaning*. Based on the *symbol grounding problem* (Harnad 1990) that states that symbols in symbolic approaches to artificial intelligence have no intrinsic meaning as they do not interact with the world, they introduce the *vector grounding problem* for self-supervised models. They claim that LLMs, in contrast to symbolic systems, do in fact have meaningful internal representations and can generate meaningful outputs. They argue that referential grounding is in fact the relevant form of grounding as it underlies all other forms of grounding (relational, communicative, epistemic and sensorimotor grounding), but that it is possible for LLMs to acquire. As the models are trained on data that is shaped by human interactions with their environment, they can develop a representation similar to the one humans who do not have access to a referent can acquire, mediated by the human producers of the data. Moreover, for models trained with RLHF, there is an additional objective that (among other goals) rewards truthfulness and factual correctness with respect to the real world.

In any way, the distributional approach is the currently most widespread approach, and often successful from a pragmatic perspective. Bengio et al. (2013) argue that a good representation is one that is useful in its downstream applications. This is certainly the case for LLMs, and it is the reason why while LLMs may not acquire understanding of the text they create, they are widely used in practice.

3 Interpretation

In Chapter 2, I emphasised the self-supervised nature of the training of neural language representations, particularly of LLMs. These models are not explicitly constructed with a defined feature set but their qualities emerge from raw text (Bommasani et al. 2021). LLMs are too complex to enable the user to comprehend their decision-making criteria and rationale, due to the sheer amount of learnable parameters and the non-linear nature of the functions they model (Mittelstadt et al. 2019). As neural language representations dispense with the need for manually crafted features, it appears natural to ask if the representations have implicitly learned similar features, even though trained on a language modelling objective only. The massive performance gain on many tasks also raises the question of what additional information causes the gain.

I adopt the definition of *interpretability* in machine learning by Roscher et al. (2020): the human understandability of *some of* the internal properties of a model (Roscher et al. 2020). Interpretability research in NLP and other machine learning-powered applications aims at creating methods and models that make machine learning systems (more) comprehensible to humans, and that test the models' behaviour under specific circumstances. Alain and Bengio (2017) argue that probing helps to *get a sense of how the training is progressing in a well-behaved model*. Inspired by such an improved understanding, interpretation studies could motivate developments of the architecture or the learning objectives that lead to better models in general or models that are better adapted for specific use cases. In the case of NLP, interpretability methods often probe for specific syntactic or semantic features that researchers hypothesize are the basis of completing a task (or a family of tasks).

This chapter is limited to post-hoc probes that are aimed at detecting linguistic knowledge in pre-trained word representations. While the community has developed a large family of interpretability methods, I will introduce four widely used categories to give a taste of the field. The first one, behavioural or zero-shot probing, provides tactical inputs to the model and inspects its corresponding output probabilities. The latter three, structural probes, mechanistic interpretability and probing classifiers, operate on the hidden representations of the model to get insights into the model's inner workings. I will summarise existing research and discuss each method's benefits and drawbacks. In Section 3.1 I will introduce behavioural probes that operate within the LLMs' training objective and compare probabilities for different queries. In Section 3.2, I introduce structural probes that investigate the geometric properties of the internal representations. In section 3.3, the field of mechanistic interpretation aimed at reverse-engineering the model to understand at the neuron level what the model is doing. Finally, in Section 3.4, I introduce classifier probes. As classifier probes are the method that our work in the interpretability field is about, I will introduce them in more detail, providing a broader overview of the concept and relevant works and discussions.

3.1 Behavioural Probes

The simplest setup for probes is the zero-shot setup, where the model's native training objective is used to query the model. As the model is left unchanged while targeted inputs are being processed and the outputs are used for the analysis, probes in this setup are commonly called *behavioural probes*. In the case of an autoregressive language model, the most common setup is the extraction of probabilities for different predictions for the next token. In the masked language modelling objective, the tokens of interest are masked and the probabilities for different predictions are compared.

Groundbreaking works in the field have been done by Linzen et al. (2016), Marvin and Linzen (2018) and Goldberg (2019). Those papers probe the syntactic abilities of language models, for example, linguistic capabilities like subject-verb agreement. A simple example of such a subject-verb agreement probe by Marvin and Linzen (2018) is the sentence:

> *The author laughs.*

The verb is masked out:

> *The author* [MASK].

The model succeeds in this example if the probability of *laughs* is higher than the probability of *laugh*:

$$P(laughs) > P(laugh).$$

The same setup has been applied to semantic tasks as well: Talmor et al. (2020) create a set of tasks that require commonsense reasoning abilities called *oLMpics*. They test various abilities of the models such as age comparison and object comparison. An example of the latter is in the following sentence:

A cat is [MASK] than a mouse.

We expect the probability of the token *larger* to be higher than the probability of *smaller*:

$$P(larger) > P(smaller).$$

The advantage of zero-shot probes is their simplicity: No training is needed, we can test the model directly without any modifications. Besides being easy to use, the fact that they are a direct probe of the model, without additional learning parameters that could influence the outcome, reduces the risk of the probe being influenced by external factors: The ability that we test for cannot be learned at probing time. The only model-external factor that influences the results is the choice of probing data, which cannot be avoided in any known setup.

A downside is that zero-shot probing is limited to certain tasks that can fit into the language modelling objective and where a comparison of output probabilities can be interpreted in a meaningful way. And even though no learning is happening in zero-shot probes, the design of the prompts affects the outcome. Success in such probes is context-dependent; the models often fail where discrepancy from their training distribution becomes too large (Talmor et al. 2020). This can make it easy to draw conclusions (both positive and negative) that do not generalize to other experimental designs.

3.2 Structural Probes

Another line of probing assumes the geometric properties of word representation vectors to be meaningful as they reflect similarities of the words' usage in the training data. It is the first of three interpretability methods that attempt to understand the vector space formed by model activations.

A classical example of structural probes are word analogy tasks such as the widely known work by Mikolov et al. on their word2vec model (Mikolov et al. 2013a,b). They show that simple addition and subtraction of word vectors can (sometimes) give meaningful results, such as in their famous `king:queen` example: When subtracting the vector for *man* from the word for *king*, and adding the vector for *woman*, the closest word vector to the resulting vector is the one for *queen*.

```
vector("king") - vector("man") + vector("woman")
≈ vector("queen").
```

The authors create a data set of various kinds of semantic and syntactic analogies called *Semantic-Syntactic Word Relationship*. It includes quadruples from domains such as country and capital, adjective and adverb, and singular and plural. They assume that a good word representation model should perform well on this data set. This assumption has been disputed, most influentially by Rogers et al. (2017) who argue that assuming linguistic relations to expose such strong regularities is psychologically not plausible: Semantic features are graded and messy, and even humans struggle with

Figure 3.1: Syntax tree recovered from BERT representations with a structural probe. Example by Hewitt and Manning (2019).

analogy tasks due to their ambiguity. And while analogical reasoning is fundamental to humans, analogies cannot be represented as binary inference rules. Rogers et al. (2017) also show experimentally that the word analogies only work when the source and target vectors are close to each other. Other studies (Baroni et al. 2014; Levy and Goldberg 2014) show that the relational similarities are not exclusive to neural word embeddings but can also occur when using count-based word embedding strategies.

Structural probes looking for sentence-level features in contextual word representations are arguably less famous than relational probes on word embeddings but there are some influential works. Hewitt and Manning (2019) develop a structural probe that tests if entire syntax trees can be extracted as a (learned) linear transformation from ELMo and BERT word representations. They assume that the number of edges between words (the depth of the parse tree) may be encoded in the representation as an L2 norm, finding out that it is indeed to some extent a structural property of the word representation space. An example for such a tree from their work is presented in Figure 3.1.

Ethayarajh (2019) investigates embeddings of BERT, ELMo, and GPT-2 with respect to how contextual they are. The author tests (among some other things) how similar the representations of the same word are in different contexts, and finds that in the upper layers, the cosine similarity decreases, suggesting that the upper layers are more task-specific. Kornblith et al. (2019) measure representational similarity between NNs trained from scratch on image classification datasets with a measure that is invariant to invertible linear transformations. They find that the representations of different datasets are similar in early, but not in higher layers.

It is an intuitive idea that the positioning of representations in the vector space, shaped by statistical patterns from the training corpus, is meaningful. However, structural probes have the practical limitation that they require strong assumptions on how the properties of interest are encoded in the representation. Those assumptions are only possible to make for certain properties that can be translated into a geometric relation.

3.3 Mechanistic Interpretation

Closely related to the structural probes, mechanistic interpretability aims at a fine-grained understanding of models at the neuron level. This field builds on the assumption that it is possible to *reverse engineer* a model from its internal representations to human-understandable algorithms, analogously to the reverse engineering of a binary computer program. The discovered algorithms are called circuits, and are subgraphs of the network that connect features in an interpretable way (Cammarata et al. 2020; Olah et al. 2020).

A disputed assumption is that the features and circuits are universal, i.e. that analogous variants exist across similar models. Some works suggest that at least otherwise equal models with different seeds converge at similar (but not equal) feature representations (Y. Li et al. 2015).

In NLP, the mechanistic interpretability work is so far centred around discovering *features*. In an early work, Karpathy et al. (2015) find several interpretable memory cells in LSTMs via activation statistics, encoding patterns like line length counters and cells that activate within brackets and quotes in code. Dalvi et al. (2019) find an individual neuron that activates at month names and one that activates at negation words, with the ten top words that activate at them being, respectively:

> **Month neuron:** *August, July, January, September, October, presidential, April, May, February, December.*
> **Negation neuron:** *no, No, not, nothing, nor, neither, or, none, (Negation) whether, appeal.*

Stanczak et al. (2022) probe for the cross-lingual overlap of neurons for morphosyntactic properties such as gender, number and tense, finding that it is significant. They conclude that this result indicates a language-independent representation of these features. The ROME method by Meng et al. (2022) employs *causal tracing* to locate factual knowledge, and to edit it to alter the GPT-2's output. The authors provide an example where the input text:

> *The Space Needle is located in the city of Seattle.*

The aim of their paper is to locate where in the model the fact that the Space Needle is in Seattle is represented, and to edit it so that the model places the needle in a different city. Their findings using ROME suggest that factual knowledge is concentrated in mid-layer feed-forward modules.

While mechanistic interpretability, popularised by the AI company Anthropic with accessible and engaging blog posts (Cammarata et al. 2020; Olah et al. 2020), has become the best-known term, attempts to discover individual neurons that represent specific properties have taken various names. For example, Dalvi et al. (2019) call their technique with this goal *linguistic correlation analysis*. Torroba Hennigen et al. (2020) name attempts to discover such neurons *intrinsic probing*, and their specific technique *dimension selection*.

Bricken et al. (2023) argue that, rather than individual neurons, linear combinations of neurons should be viewed as the features and building blocks of mechanistic interpretability. This is because individual neurons have a property that is called polysemanticity: They activate at inputs that do not correlate in a human-understandable way. The work uses a sparse autoencoder setup on a one-layer transformer language model, and finds *relatively* monosemantic (and thereby interpretable) features. This is not a new but rather a rebranded insight: Combinations or rankings of multiple neurons have also been the building blocks of the analogous techniques used in academic research (Dalvi et al. 2019; Stańczak et al. 2023; Torroba Hennigen et al. 2020).

Fully understandable circuits have, not surprisingly, so far only been discovered for very well-defined tasks. Nanda et al. (2023) study one-layer transformers for modular addition. They are able to fully reverse-engineer the model and find that it performs the task by mapping the inputs to a circle which it uses to combine the inputs with its feed-forward and attention layers. If similar results are possible with much deeper models, and especially if such interpretable circuits will be possible for real-world applications given the noisy nature of natural language and many NLP tasks remains to be seen.

3.4 Probing Classifiers

The way to investigate representations that I focus on in this thesis is through the use of supervised probing classifiers (Alain and Bengio 2017; Hupkes et al. 2017). Probing classifiers are simple classifiers that are trained to solve diagnostic prediction tasks considered to require relevant linguistic information, such as parts of speech, syntactic structure, or semantic roles, from frozen model parameters. If the classifier is capable of predicting the property of interest to a high degree, it is concluded that the property (or, more precisely, information highly correlated to the property) has been learned implicitly by the model.

Probing classifiers are often used to measure quantities of interest in different parts of a model, to see where in a model a specific property is best found. Alain and Bengio (2017) compare probing to using thermometers to measure the temperature simultaneously at different locations within the model.

3.4.1 Definition

A probing classifier is a supervised classifier, typically a simple feed-forward network or a linear network, that is trained on a probing task (diagnostic task). It uses datasets $D = (x_i, y_i)_i$, where each x_i is the continuous vector representation of a neural language model at some specific layer and y_i is the gold-standard label for the probing task. The x_i are often word-level representations as typical probing tasks are often at the word level, such as parts of speech, syntactic dependencies, semantic roles and co-referent entity mentions. The classifier's performance is often measured with classification metrics such as accuracy and can be compared to baselines or across layers.

To give a simple example, the dataset D may contain sentences where each word is annotated with a part of speech tag, such as *noun*, *verb* or *adjective*, which become the y_i. If we feed the sentence through an LLM (e.g. BERT), we can extract each token's internal representation from a specific layer (e.g. layer 3). As the words may be split into several tokens by BERT's tokeniser, we may need to realign them, e.g. by adding up the representations of the tokens that the word was split into. Now that we have one representation for one word, we also have our x_i. The dataset D can now be used to train and evaluate the probing classifier and see how much useful information for part-of-speech tagging BERT contains at layer 3. By building analogous datasets for the other layers, we can compare the probe's performance across layers and see where in the model most useful information is found.

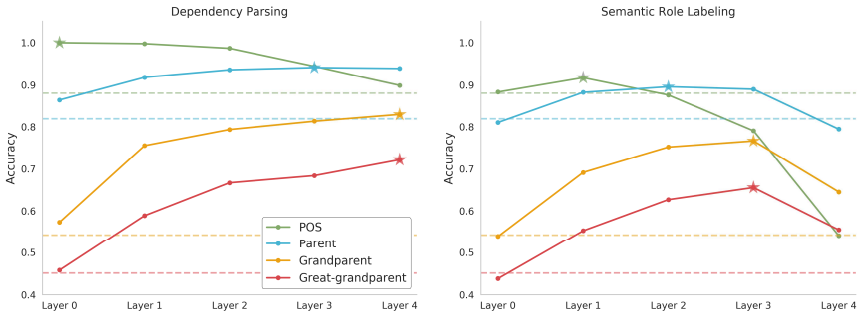


Figure 3.2: Results from Blevins et al. (2018), with a probe on a dependency paring and a semantic role labelling model that exposes a hierarchy of tasks: Part-of-speech information peaks first, then syntactic parents, and then syntactic grand- and great-grandparents.

3.4.2 Influential Works

Hupkes et al. (2017) propose probing classifiers, or diagnostic classifiers as they call them, to investigate how RNNs capture the hierarchical compositional semantics of natural language. To isolate this mechanism from other properties of natural language (like structural and lexical ambiguity, irregular paradigms, multi-word units and idiomatic expressions), they use a synthetic dataset of nested arithmetic expressions. They find that RNNs roughly follow a cumulative strategy, as the intermediate value of the expression up to the point from which the current representation is taken can be predicted more accurately than the value within the current brackets, which would be necessary for a recursive strategy of solving the expressions.

Various papers found that information is structured hierarchically within the models. Belinkov et al. (2017a) probe machine translation models for part-of-speech and full morphological tagging. They find that the lower layers of the encoder perform better for this type of information, while deeper layers have less information about it. As deeper networks are however crucial for translation quality, they hypothesize that deeper layers are specialised in meaning. They also find that the decoder part of the network contains little morphological information. In a subsequent work, Belinkov et al. (2017b) find that indeed, the higher layers contain more semantic tag information. Blevins et al. (2018) probe four different LSTM-based models: a syntactic dependency parser, a semantic role labelling model, a machine translation model and a language model. They discover a soft hierarchy within the models, with part-of-speech information more prevalent in the lower layers, followed by syntactic parents, grandparents, and finally great-grandparents. Two example results for models they examine are presented in Figure 3.2.

In the paper that introduces the ELMo model, Peters et al. (2018) find that part-of-speech tags are better predicted from the first hidden layer, while word sense information is more prevalent in the second layer. In a comprehensive probing study with many different tasks, Tenney et al. (2019) discover that in the BERT model, the tasks are ordered analogously to a hand-crafted NLP pipeline from pre-neural times:

from POS tags over syntactic dependencies, named entities and semantic roles to coreference information. Y. Lin et al. (2019) use probing classifiers for tasks that require different sorts of syntactic information, from linear information like a word's position to hierarchical information like the main auxiliary or the subject noun of a sentence. Their results indicate that embeddings from lower layers contain more linear information while higher layers contain more complex hierarchical features. That semantics can be found in the higher layer while the lower layers contain more syntactic information is also reported by Raganato and Tiedemann (2018) who use probing classifiers on part-of-speech tagging, chunking, named entity recognition and semantic tagging on Transformer encoders for machine translation. Similarly, Jawahar et al. (2019) demonstrate that BERT learns surface features at the bottom layers, syntactic features in the middle layers, and semantic features in the top layers, suggesting that BERT requires deeper layers to learn long-distance features.

Blevins et al. (2022) analyse the pre-training dynamics of a multilingual model by probing various training checkpoints. They find that while monolingual capabilities are acquired early, cross-lingual capabilities emerge later in training, to a varying degree depending on the language pair. They also find that linguistic knowledge wanders from higher to lower layers during pre-training. K. Zhang and S. Bowman (2018) probe representations learned with different learning objectives, including language modeling and translation, for syntactic tasks, finding that the representations trained on bidirectional language modelling contain the most useful information.

Probing has also been used for the evaluation of document or sentence embeddings. Adi et al. (2017) argue that downstream task evaluation is very coarse-grained and does not tell much about the types of information contained in the embeddings, and therefore, generalisable conclusions cannot be drawn. They therefore introduce a set of three tasks that capture the most basic properties of a sentence: word order, word content, and length. Conneau et al. (2018) probe sentence embeddings for ten simple linguistic features, finding that encoders contain a wide range of linguistic information compared to baselines. They argue that for these simple tasks, it is easier to control for biases than for downstream tasks.

Slobodkin et al. (2023) probe if the embedding space of a model encodes the answerability of a question with the help of a dataset that contains both answerable and unanswerable questions. They find that a probe on the last hidden layer reaches a much higher accuracy than a probe on the first layer, and conclude that there is an answerability subspace within the representations.

3.4.3 Methodology

Several works have raised concerns that a powerful probe may simply learn the task by storing information in its own parameters, rather than exposing features from the representations. The efforts in probing were in favour of placing restrictions on the classifiers. Alain and Bengio (2017) explicitly measure the level of *linear separability* with their probes, arguing that a linear classifier avoids the problem of local minima. Hewitt and Liang (2019) restrict the model in several ways: In making them linear like Alain and Bengio (2017), but also in the hidden size and the amount of training data, arguing

that a reliable probe should be *selective*: It should perform well on the actual task, but on the same time be unable to learn (memorise) a *control task* that randomly assigns word forms labels. Voita and Titov (2020) propose to tackle this problem by using the minimum description length (Rissanen 1978), an information-theoretic estimate of complexity that takes into account the complexity of the model (its codelength). Pimentel et al. (2020a) explicitly take into account the accuracy–complexity trade-off. They argue that optimising for performance alone does not allow for conclusions about the representations, but that optimising for complexity alone will make the probe unable to exploit complex information about a word's identity. Hence, a probe should be Pareto optimal: In a family of probes, there should be no probe that has a higher accuracy and at the same time a lower complexity.

On the other hand, Pimentel et al. (2020b) argue that a probe should be as expressive as possible. In their information-theoretic framework, they measure the mutual information between a representation and a linguistic property (the labels) and argue that it is best measured with a powerful model. Saphra and Lopez (2019) argue that the intermediate representations of neural networks cannot be expected to be linearly separable, and that the non-accessibility of a property to a linear probe cannot be expected to imply that this property does not exist in the representation.

Another concern about standard probing methodologies is that they do not take into account if the detected information in the representation is actually relevant to the model when performing language modelling or downstream tasks. Ravichander et al. (2021) show that linguistic properties can be encoded in the representations of a model (in their case, a BiLSTM) even if they are not relevant to the models' training task. Based on this problem, Elazar et al. (2021) propose *amnesic probing*, a technique that shifts the focus from the question of whether certain information is encoded in the representation to the question of how this information is being used by the model. They use the Iterative Nullspace Projection algorithm (Ravfogel et al. 2020) to delete linear information associated with a task, and measure the effects of the intervention on the language modelling performance. The results indicate that not all tasks that are easily learned by a probing classifier influence the performance: While the impact low-level syntactic information is the strongest, named entity information has a small and phrase detection information no impact. Rozanova et al. (2023) build on this work and propose *mnestic probing*, a method that reverses amnesic probing and keeps only the information that has been deemed task-relevant. They show that this strategy is more informative for their target task, natural language inference, where high-dimensional representations meet a small set of class labels.

3.4.4 Comparison to other Methods

Compared to the analysis of attention modules, probing classifiers have the advantage that they are largely agnostic to the architecture of the neural network, as they only rely on the parameters of the feed-forward part. The latter can however also be seen as a limitation, as probing classifiers ignore the specific function of other parts of the architecture. In particular, they do not consider the function of the attention heads, which have been shown to exhibit interpretable meanings (K. Clark et al. 2019). However, the *effect* of the attention heads should be observable from the representation

in the feed-forward part. And in fact, studies on attention and syntactic structure show similar results to probing: Vig and Belinkov (2019) show that the attention heads in the highest layers capture the most distant syntactic dependency relationships, while most dependency relationships are to be found in the middle layers.

Probing classifiers may also be applicable to a broader set of tasks, as there may not be a suitable attention head that covers a sufficiently similar structure for every task. Some tasks may depend on information aggregated from various attention heads in a way that is not easily observable. The broader applicability shows even more for the comparison to structural probes and behavioural probes. Those two techniques are, as discussed earlier, only applicable to tasks with very specific properties. For probing classifiers, the only limiting factor is that we need annotated data that can be aligned with the representations. Feature discovery techniques in mechanistic interpretability are however closely related to probing classifiers. Consequently, they can be applied on the same set of problems, and give us similar types of insight (and more, as it is possible to obtain results on a more fine-grained level).

A disadvantage compared to behavioural probes is that for probing classifiers, we need parameter access to the model. There is no way to analyse models with probing classifiers, structural probes or mechanistic interpretability if we access them as a blackbox via an API. On the other hand, behavioural probes do not allow comparison of layers and thereby cannot offer any insights about models' internal dynamics.

As outlined earlier in this section, what can be concluded from the results of a probing classifier is subject to discussion. The usage of a trainable classifier makes the methodology more indirect than other methods that work with the model or representation as-is. This explains why there is more methodological discussion around probing classifiers than around most other probing methods. This is the motivation behind our work on probing classifiers: How do we know what we can conclude from a probe? And how can we make the method as reliable and insightful as possible? We approach these questions in Papers I, II and III.

4 Explanations

Complementary to analyzing the properties of a model globally as described in Chapter 3, it is insightful to understand the reasons and mechanisms behind *individual predictions*. This can give both developers and users crucial context in order to know if predictions are reliable, or if a model uses undesirable shortcuts and biases.

The question what constitutes an explanation touches philosophy and the social sciences. In social science literature, explanations have been viewed as an answer to *what*–questions, *how*–questions, and *why*–questions (Miller 2019). For explainable AI, explanations are likely to answer *why*-questions: Why is A the answer to question or query Q? Or, more specifically: Why does the model infer answer A? By answering this question, explanation methods aim at making the outputs of the NLP model and its behaviour more intelligible to humans. The explanation mechanisms can either be part of the model itself, or they can have access to the model parameters, or they can simulate its internal processes by collecting outputs for targeted inputs.

The currently most widespread approach in explainable NLP is *attribution methods* (Section 4.1), which identify those components of the model input that have the most impact on the output. However, the applicability and clarity of these methods have been criticised as limited (Bansal et al. 2021; Papenmeier et al. 2022; Wiegreffe et al. 2021). In addition, humans prefer explanations that target the mechanism of *how* the model arrived at a conclusion over explanations that solely list covariant factors (Lombrozo 2006). An alternative that focuses on both covariant factors and the mechanism are *procedural* explanations (Section 4.2) that offer a complete reasoning path from the input to the prediction, but that have an even more limited applicability. As a third alternative, free-text explanations (Section 4.3) have recently gotten traction

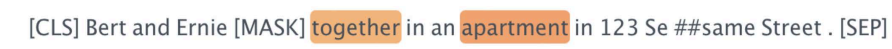


Figure 4.1: An example for a feature attribution-based explanation in the masked language modelling task, highlighting the most relevant tokens for predicting the masked-out token. The missing word in this example is *live*. Created with AllenNLP Interpret (E. Wallace et al. 2019b).

in NLP (Marasovic et al. 2022; Wiegreffe et al. 2022). They are easily accessible to users, as they imitate human explanations, and are flexible in the tasks they can be used for and the types of reasoning they can express. However, there are also significant concerns. One is that free-text explanations, like human explanations, focus on *proximal mechanisms* rather than complete sets of reasons or complete logical deductions (Tan 2022). Explaining a model in this way can seem unintuitive as from a technical system, many people expect the ability to explain things in a technical, not a social form. The most widespread concern is however that the connection to the model's reasoning process is unclear (Bommasani et al. 2021). This is also the case for many other explanation techniques (as I will outline in this chapter), but even more evident for explanations that are largely generated by the same LLM that generates the predictions, with the same mechanism, but no easy way to reliably test the connection between answer prediction and explanation generation.

The object of study of our own work has been natural language explanations, which I will therefore discuss most extensively in Section 4.3. To set natural language explanations in context in the field of explainable NLP, I will however start with introducing two other impactful approaches: input relevance measures in Section 4.1 and procedural explanations in Section 4.2, with a focus on their limitations that make natural language explanations stand out.

4.1 Input Relevance Measurements

Input Relevance Measurements, also called feature attribution techniques, explain model predictions by identifying critical parts of the input. For a document classification task, this could be a collection of words that were particularly important for the assignment of a document to a certain category. Wiegreffe and Marasovic (2021) refer to explanations produced by feature attribution as *highlights*, Tan (2022) calls them *evidence*. They are the most common form of explanation for explainable NLP. An example for a feature attribution-based explanation can be found in Figure 4.1.

The idea behind feature attribution is that explanations have classically been defined as sets of *causes*, that is, empirical conditions or natural laws that lead to the explanandum—the statement about the phenomenon in need of explanation (Lombrozo 2006; Moravcsik 1974). However, Lombrozo (2006) argues that humans prefer explanations that cover the *mechanism* of how causes lead to an outcome. Attribution leaves out all this information on *how* the highlighted information is being used by the model (Rudin 2019). As such, it does not map to any form of everyday explanation, and Tan (2022) even argues that stand-alone evidence is not a form of explanation.

Instead, we may think of highlighted parts as *causes* that have the *effect* of making the model produce a certain output (Keil 2006).

When it comes to the utility of attribution methods, studies have arrived at mixed conclusions. Bansal et al. (2021) test if explanations increase the performance of human-AI teams and find that they do not provide more value than confidence scores alone: they yield an increase in accuracy when the model is correct, but a decrease when it is wrong. Papenmeier et al. (2022) report the results of a large user study where faithful explanations did not increase user trust or understanding compared to random or no explanations. Chu et al. (2020) show that for image classification, accuracy, trust, and understanding are not significantly improved by providing visual explanations. However, the utility of explanations may vary across tasks and designs. In particular, they may be helpful for tasks that require navigating external information sources. For example, González et al. (2021) retrieve answers for challenging open"-domain questions from corpora and find that explanations help users predict errors better than confidence only. Shen et al. (2022) show that the length of the explanation is crucial for understanding.

Apart from its restriction to information in the model input, a further shortcoming of feature attribution is its limitation to tasks where either the input itself or the retrieved document explicitly contains all relevant context. More open-ended tasks that rely extensively on information inferred from the model parameters or other knowledge and reasoning, such as open-ended or multiple choice question answering (Talmor et al. 2019) or algebraic word problems (Ling et al. 2017), cannot be satisfactorily explained using feature attribution.

4.1.1 Attention

An intriguing technique for assessing feature relevance, as it is a pre-existing component of many NLP models including Transformers, is looking at the *attention scores*. However, it is subject to discussion how meaningful the scores are as the models are not specifically trained on attending to relevant inputs, let alone to rank inputs in attention scores.

Serrano and Smith (2019) investigate the input and output of an attention layer before and after an intervention on the attention weights in order to assess if attention distributions can be used to identify the most relevant information. They perform a number of experiments in a text classification system in its final attention layer to judge attention's capability of being used as an importance ranking. In the first experiment, they remove the attention weight of the component with the highest attention and compare the effect with a randomly removed attention weight. In this case, the conclusion is tentatively positive: The highest-scored element has a much larger effect when being removed than the random element. In another experiment, they erase attention weights subsequently starting at the top of the ranking and observe at which point the model's decision changes. They compare the attention distribution with three other rankings: a random order, an order determined by the gradient of the decision function with respect to each attention weight, and the attention weights supplemented with information about the gradient. They find out that while the

attention weights distribution flips the decision faster than the random order, the two latter approaches flip faster than the (pure) attention weights, suggesting that attention weights are not an optimal predictor for models' decisions although they do sometimes correlate with importance, but in a noisy way.

The paper *Attention is not Explanation* (Jain and B. C. Wallace 2019) also explores how well attention weights can capture the relative importance of an input, but in the context of the full model instead of the final attention layer as done by Serrano and Smith (2019). Based on experiments with comparisons of the attention distributions to gradient-based feature importance measures and an evaluation with adversarial attention, Jain and B. C. Wallace (2019) claim that attention weights are weak predictors. This paper has been criticised in its methodology and conclusions by Wiegreffe and Pinter (2019) whose answer paper is called *Attention is not not Explanation*. In particular, Wiegreffe and Pinter (2019) question the adversarial attention experiments as they do not take into account that attention weights are not trained independently but end-to-end with the full model, and as the manipulated attention scores are not truly adversarial. Wiegreffe and Pinter (2019) provide alternative approaches for testing the meaningfulness of attention distributions, in particular by training examples that they show to be truly adversarial. Based on the poor performance on those new adversarial examples, they conclude that attention scores *can* provide coherent interpretations.

It is apparent that researchers have different perceptions of terms like explainability and interpretability. This is also reflected in the definitions that they give. Serrano and Smith (2019) write that "*[i]n order for a model to be interpretable, it must not only suggest explanations that make sense to people, but also ensure that those explanations accurately represent the true reasons for the model's decision.*" Jain and B. C. Wallace (2019) state that their question is "*[. . .] whether attention suffices as a holistic explanation for a model's decision"*. Wiegreffe and Pinter (2019), on the other hand, claim that attention is usually treated as *an* explanation, not *the* explanation, and that most earlier work claimed attention rather to be "*providing plausible rationales*" than guaranteed faithful explanations, so that their interpretations of the attention models can still be seen as valid.

While the previously discussed papers studied attention mechanisms in models with RNNs, others have investigated attention distributions, but to self-attentive Transformer models with multiple heads. Voita et al. (2019) look at distributions in the individual attention heads of a transformer model for machine translation, finding that there are heads that take interpretable functions such as positional heads attending to a specific relative position, attention heads that model specific syntactic relations and a head attending to rarest tokens in a sentence. In their evaluation of a pruning experiment, the authors show that the interpretable heads are more relevant for the system performance than the others, and that most heads, especially in the encoder, can be pruned without seriously affecting the system's performance. A related method is used byK. Clark et al. (2019) who analyze BERT's attention heads and also find several functional heads that attend for example to the direct objects of verbs, determiners of nouns, objects of prepositions, and coreferent entity mentions.

4.1.2 Other Attribution Methods

Besides attention, there are more methods to extract input relevance scores. There are two main approaches: perturbation-based surrogate models, and gradient-based methods.

Surrogate models are smaller, more interpretable models that approximate model decisions. They are model-agnostic, meaning that they treat the model they are supposed to explain as a blackbox and query it using small perturbations around the input data to infer the relevance of each input feature. Examples for such models are Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro et al. (2016)) and SHAP (Lundberg and S.-I. Lee 2017). While surrogate models are widely used due to their simplicity and wide applicability, they have been shown to be unstable for non-linear models (Alvarez-Melis and Jaakkola 2018; E. Lee et al. 2019), and the sampling procedures to perturb the input can lead to uncertainty (Yujia Zhang et al. 2019). Moreover, surrogate models are sensitive to adversarial attacks: They can even be manipulated to not reflect the true features and biases underlying the decision-making (Dombrowski et al. 2019; Slack et al. 2020).

Another family of techniques that is not strictly model-agnostic but can be used for all neural network models uses gradients to compute the contribution of input features to the model's decision (J. Li et al. 2016; Simonyan et al. 2014; Sundararajan et al. 2017). As gradient-based models use model parameters only, requiring just a backward pass without modifications, they are considered more faithful than surrogate models. Faithfulness is particularly relevant given that the intended user of the generated score is often the model developer (Bastings and Filippova 2020). For other target users however, it is relevant that even the gradients of a model can be subject to adversarial attacks: J. Wang et al. (2020) train models whose gradients place high attribution to tokens not relavant to the task without affecting the predictions.

4.1.3 Evaluation

As outlined in this section, no input relevance measure is fully reliable and robust. Moreover, different techniques can produce contradictionary results (Atanasova et al. 2020a). To systematically compare and evaluate them, various criteria and measures have been proposed.

Alvarez-Melis and Jaakkola (2018) propose measures for the robustness of the explanations to slight variations of the input. They find that while no method they test is robust, gradient-based methods perform better than perturbation-based methods. Yu et al. (2019) suggest a set of three properties that a rationale should maximise: *sufficiency*, meaning that it is possible to make the prediction with the selected features only, *comprehensiveness*, meaning that the explanation includes *all* relevant features, and *compactness*, meaning that the explanation should not be longer than necessary and that it should be continuous. Atanasova et al. (2020a) propose a more extensive set of criteria, including the agreement with human-annotated rationales, confidence measures, faithfulness (as measured by performance difference when leaving out the most salient tokens), and consistency. Even they find that gradient-based methods are fulfilling their criteria best.

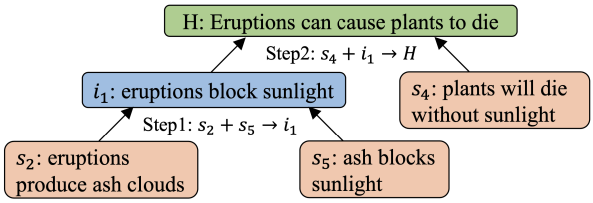


Figure 4.2: Example for a deductive explanation: A constrained METGEN (Hong et al. 2022) tree for science question answering. The question in this example was: *How might eruptions affect plants?*, the answer, as shown in green in the figure: *Eruptions can cause plants to die*. Orange denotes facts; blue intermediate conclusions. Figure adapted from Hong et al. (2022).

For the comparison with human rationales, there exist datasets with annotations for relevant input spans. The benchmark ERASER (DeYoung et al. 2020) collects seven such datasets. The authors of ERASER propose to measure agreement with these human annotations, but also faithfulness scores based in sufficiency and comprehensiveness as the goal of outputting attribution scores is to create not only *plausible* but also *faithful* explanations. The importance of the disentanglement between the two has been noted by Jacovi and Goldberg (2020) who point out that human ratings or gold standards are inappropriate for faithfulness evaluation as plausibility to humans does not indicate what a machine learning model is doing internally.

4.2 Deductive Procedure

Deductive procedures are grounded in the input and provide step-by-step rules that lead to the prediction. They are less common because they are only applicable to a small share of NLP tasks (Tan 2022) but provide complete inference chains where all intermediate steps can be checked. Long before deep learning became popular, procedural explanations have been used in artificial intelligence to learn generalization by capturing the structural relationship of a problem (DeJong and Mooney 1986; Lewis 1988; Mitchell et al. 1986).

Narayanan et al. (2018) employ explanations in the form of *decision sets*, mappings of inputs to outputs via a set of rules. In user studies, they search for explanations that humans can utilize best, finding that more complex explanations are harder to process and less satisfactory. Hong et al. (2022) build constrained trees consisting of entailment steps for science question answering. An example for such a tree generated with their METGEN system can be found in Figure 4.2. Ling et al. (2017) and Jie et al. (2022) generate the intermediate steps necessary to solve math word problems. This is similar to the recently famous Chain-of-Thought (CoT) generation (Wei et al. 2022), where the model generates intermediate reasoning steps prior to the prediction in a zero-shot setting. However, in CoT, the completeness and correctness of the intermediate steps is neither enforced nor typically evaluated; its main goal is the improvement of the prediction accuracy.

There are two main limiting factors in deductive procedures. One is applicability: The prediction problem needs to be fully formalisable, which is a very strong assumption as most dynamical systems are not characterisable as (interpretable) computations (Cummins 2000). Also, procedures are limited to problems solvable by simple and transparent algorithms, such as decision trees. Humans, on the other hand, explain at different levels of abstraction, even if their understanding is coarse and fragmentary (Keil 2006). To understand how rare purely formal reasoning is in humans, we should consider that even discovery-focused parts of mathematics have informal components: They require the refinement of guesses by speculation and criticism, heuristics, and exploration (Lakatos 1963).

The second limitation of deductive procedures is their understandability: If the explanation exceeds a certain length, it will be hard for a human user to follow. The causal and relational complexity of the real world would however require deductive procedures of unbound length, making understandable procedures utopian.

4.3 Natural Language Explanations

Generating natural language explanations has gained relatively little attention until a few years ago. While some works use more restrictive techniques to generate textual explanations, such as template-based approaches (N. Wang et al. 2018; Yongfeng Zhang et al. 2014) or approaches based on extractive summarisation (Atanasova et al. 2020b), it was only with the emergence of GPT-2 and other generative Transformer models that they gained more traction. Datasets with human-written free-text explanations[1] for the correct label were created for tasks like natural language inference (Camburu et al. 2018) and multiple-choice question answering (Aggarwal et al. 2021; Rajani et al. 2019), along with models fine-tuned to imitate these explanations.

Natural language explanations address many of the limitations that attribution methods and procedural methods have. They are easily accessible to human users than other forms of explanation, especially to end users without a technical background. Moreover, they can incorporate forms of reasoning not covered by the other methods: They allow for the inclusion of any input-external knowledge and any type of reasoning that can be expressed in natural language, while not being restricted to tasks where it is possible to provide a complete reasoning path.

The incorporation of free-text explanations in the learning process has in some cases lead to an increase in performance and robustness. The idea behind this approach is that with the explanations as additional supervision, models can be guided to make decisions in a more accurate way and rely less on spurious correlations in the dataset, as the decision-making process is better aligned with the expected forms of reasoning. This may also improve the model's robustness, as it may reduce the dependence on cues from one dataset that do not generalise to other datasets. While similar ideas

[1] At this point, I use *free-text explanations* as a narrower term than *natural language explanations* as it excludes template-based or extractive natural language explanations. In the remainder of this chapter, I use the two terms interchangeably, both referring to explanations generated token-by-token by an LLM.

are also applied to attribution methods (Chen et al. 2022), the freer form of natural language explanations may extend the benefits to more tasks.

The need of manual annotations of explanations, the creation of which is costly, is a limiting factor in generating free-text explanations (Belinkov and Glass 2019). This problem may have been partially resolved with LLMs like GPT-3 and beyond that have few-shot and zero-shot capabilities. An objection to free-text explanations that is of enduring relevance is however that the explanations have no obvious connection to the label prediction. As Bommasani et al. (2021) write:

> "[...] *there are reasons to be skeptical: language models, and now foundation models, are exceptional at producing fluent, seemingly plausible content without any grounding in truth. Simple self-generated "explanations" could follow suit. It is thus important to be discerning of the difference between the ability of a model to create plausible-sounding explanations and providing true insights into its behavior.*"

Bommasani et al. (2021)'s sceptical remarks are justified: We cannot trust LLM-generated explanations to accurately present its inner workings when generating a label. However, as I will discuss in this section, just as most other explanation techniques, they are able to provide us with *some* insight about the model's inner workings. As the kind of insight they provide us with are not currently achievable by other techniques, they are a valuable component of model explainability and can be useful to both developers and users.

In this section, I introduce influential applications and datasets (§4.3.1) and common approaches to generating free-text explanations (§4.3.2). Finally, I discuss the evaluation of free-text explanations and the question how we can determine if they are faithful to the prediction process (§4.3.3).

4.3.1 Applications and Datasets

Datasets containing free-text explanations broadly fit into two categories: Tasks with a focus on logical and mathematical reasoning, and tasks that require factual knowledge and commonsense reasoning.

Logical Reasoning

The first category of datasets require logical or arithmetic reasoning between statements to solve a task.

A widely used dataset that pioneered the generation of free-text explanations is eSNLI by Camburu et al. (2018). They add explanations to the natural language inference (NLI) datasets SNLI (S. R. Bowman et al. 2015). The task of SNLI is to classify the entailment relation of two sentences (the *premise* and the *hypothesis*) into *contradiction*, *neutral* or *entailment*. An example of this is the following, where the explanation points at the logical problem in inferring the hypothesis from the premise, resulting in the label *contradiction*:

> **Premise**: *A shirtless man is singing into a microphone while a woman next to him plays an accordion.*
> **Hypothesis**: *He is playing a saxophone.*
> **Label**: *contradiction.*
> **e-SNLI Explanation**: *A person cannot be singing and playing a saxophone simultaneously.*

The SNLI dataset has been shown to contain unintended cues that models can rely on when making predictions (Gururangan et al. 2018). Other, more diverse datasets also rely on such cues (T. McCoy et al. 2019), making it uncertain how well models solve the actual task even if the model scores high in an in-domain evaluation. Textual explanations that exhibit the expected reasoning can be valuable to assess the potential of the model to solve the task accurately and its limitations. In this line, Zhou and Tan (2021) extend the adversarial HANS dataset (T. McCoy et al. 2019) with explanations and show that model-generated explanations contain a high lexical overlap with the human-written explanations but hallucinate information and miss crucial relations.

Guiding models through explanation generation sometimes improves their performance. In chain-of-thought prompting (Wei et al. 2022), the model generates intermediate reasoning steps before the prediction in a few-shot setting on arithmetic word problems. The approach has been shown to substantially boost the performance for appropriate tasks when combined with very large models. The most widely used dataset for chain-of-thought prompting is GSM8K (Cobbe et al. 2021). GSM8K consists of a question, the reasoning path, and the final answer, as shown in the example below:

> **Question**: *Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?*
> **Answer**: Weng earns 12/60 = $12/60=0.2 → 0.2 per minute.
> Working 50 minutes, she earned $0.2 x 50 = $0.2*50=10 → $10.
> **Final Answer**: $10

Kojima et al. (2022) do zero-shot prompting on a set of arithmetic reasoning tasks (including GSM8K) and other logical reasoning tasks, simply appending *Let's think step by step* to the prompt. This alone results in a substantial task performance increase.

For e-SNLI however, such increases have not been observed. Zhao and Vydiswaran (2020) address this issue and identify two problems: Firstly, unlike humans, past models did not consider alternative explanations but focus on explaining the correct label. Secondly, the models do not reason about the correctness of facts but focus on creating well-formed sentences. They include a candidate explanation for every possible label, and use an instance selector to reason about which one is correct. The performance increase they report with this approach is however slight.

However, departing from training on the e-SNLI explanations, He et al. (2023) perform few-shot prompting similar to chain-of-thought prompting on several NLI datasets including SNLI but also various harder and adversarial datasets. They show that this setup improves task performance and robustness.

Knowledge-based Reasoning

The second category of datasets containing explanations that I will present are tasks that require commonsense and factual knowledge that is not present in the input for their reasoning.

CoS-E (Rajani et al. 2019) and ECQA (Aggarwal et al. 2021) extend the multiple-choice question answering dataset CommonsenseQA (Talmor et al. 2019) with crowd-sourced explanation. While CoS-E provides a brief explanation only for the correct answer, ECQA also contains an explanation for each of the four incorrect answers, with a reason why this option is wrong. In addition, ECQA features a long-form explanation that contrasts all answer options. The following is a shortened example from ECQA, with the question, the correct answer option along with an explanation, as well as one wrong answer option along with an explanation:

> **Question:** What is something that people do early in the day?
> **Correct Answer:** Eat eggs.
> **Explanation:** People generally eat breakfast early morning. People most often eat eggs as breakfast.
> **Negative Answer (Example):** Believe in God.
> **Explanation:** Believing in God is not restricted to a specific part of a day.

In the Commonsense Validation and Explanation (ComVE; C. Wang et al. (2020)) dataset, the task is to predict and explain whether a natural language statement makes sense. In the prediction task, the model is presented with two statements, and needs to select the nonsensical one. The explanation task has two modes; a multiple-choice mode where the correct explanation must be selected from three explanations with similar wordings, and a free-text generation mode. An example for a multiple-choice instance containing a nonsensical statement and three options for explanations is the following:

> **Statement:** John put an elephant into the fridge.
> **Option 1:** An elephant is much bigger than a fridge. (correct)
> **Option 2:** Elephants are usually white while fridges are usually white. (wrong)
> **Option 3:** An elephant cannot eat a fridge. (wrong)

For a similar multiple-choice question answering dataset, Latcinnik and Berant (2020) include an intermediate textual layer in their model to generate explanations. However, they do not perform supervised training with annotated explanations as a target but train their models with weak supervision to produce explanations that are useful for the downstream classifier. Park et al. (2018) create the visual question-answering datasets VQA-X and ACT-X by enriching existing datasets with textual justifications for answers as well as relevant positions highlighted in the image. They show that it depends on the sample if visual or textual explanations are more useful, concluding that multimodal explanations are preferable for multimodal tasks.

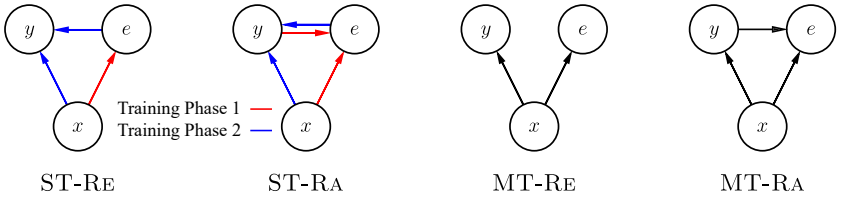


Figure 4.3: Graphical representation of the categorisation proposed by Hase et al. (2020). x is the input, y the output and e the explanation. Figure by Hase et al. (2020).

4.3.2 Approaches

In the earlier works on generating free-text explanations, the explanation is generated either completely independently of (Rajani et al. 2019) or prior to the label prediction model (Camburu et al. 2018; Latcinnik and Berant 2020). As Latcinnik and Berant (2020) acknowledge, this is a weak form of explanation as the prediction process cannot influence the explanation generation. Hase et al. (2020) introduce a categorisation of the approaches, calling the former type of approach that conditions the explanation only on inputs *reasoning* (RE) mode, and the latter that considers the input as well as the label *rationalising* (RA) mode. They also consider if the explanations are used as an additional input in a pipeline model which they call a *serial-task* (ST) approach or explanations are generated jointly with the labels, called a *multi-task* (MT) approach. A graphical representation of the four approaches is presented in Figure 4.3. Hase et al. (2020)'s multi-task reasoning mode has subsequently become known as *self-rationalising models*, the term that I adopt in this thesis.

With the increasing multi-task capabilities of LLMs, self-rationalising models have become increasingly common. Narang et al. (2020) show that they can successfully generate labels and explanations at the same time, and that the self-rationalisation capabilities can, to some extent, even be transferred to other datasets. Wiegreffe et al. (2021) show that self-rationalising models have a higher performance than serial-task architectures.

These first approaches all used fine-tuned models. After the release of GPT-3 and the increasing capabilities of models in in-context learning, Marasovic et al. (2022) were the first to propose a few-shot approach that jointly generates model predictions and explanations. They explore how the prompt should be formatted for such a setup, and show that while the prompt has a significant impact on the performance, humans rate their generated model explanations as significantly less plausible than human-written explanations.

These findings have also been exploited for the generation of datasets. Synthetic, LLM-generated datasets are becoming more common as they reach a reasonable quality for many NLP tasks and applications with a fraction of the cost, and the same appears to be true for LLM-generated natural language explanations datasets. Wiegreffe et al. (2022) generate candidate explanations with GPT-3 and train an acceptability filter based on human ratings of the explanations. They show that human raters often prefer the LLM-generated over the human-written explanations. A similar result has been reported by He et al. (2023) who compare few-shot prompting for NLI with human-annotated

versus ChatGPT-annotated explanations. They find that, surprisingly, humans prefer the LLM-annotated explanations while the human-annotated explanations are more beneficial for task performance for four out of five models.

4.3.3 Evaluation

Similar to the evaluation of input relevance measurements in Section 4.1, there are several properties of natural language explanations that can be evaluated but that should not be conflated. As we evaluate similarity of input attribution methods to human relevance judgements, we can evaluate the similarity of model-generated explanation to human-written explanations. Faithfulness evaluation is even more challenging and more experimental for free-text explanations than it is for input attribution methods, but nonetheless, there exist approaches that attempt to estimate this property.

Similarity to Human Explanations and Plausibility

Similar to other text generation tasks, the automatic evaluation of explanations has the challenge that the same thing can be expressed in many different ways. For many tasks and data instances, it is even possible to create various valid reasoning chains. Therefore, even where human-annotated explanations exist, it is not necessary to expect exactly the same explanation from the model, we can at most expect a similar meaning. Therefore, test generation metrics that measure the semantic overlap between generated and annotated explanations are commonly used.

Two widely used metrics in text generation tasks are BLEU (Papineni et al. 2002), a group of metrics originally developed for machine translation that measures the n-gram precision of generated text and ROUGE (C.-Y. Lin 2004), a group of metrics originally developed for text summarisation that measures the n-gram recall. BLEU and ROUGE however measure the surface overlap between texts and require exact matches of the n-grams. In consequence, they are not robust to lexical and syntactical variations and are less indicative for criteria such as the *acceptability* of a generated text (Ananthakrishnan et al. 2007; van der Lee et al. 2021). For this reason, the BERTScore metric family (T. Zhang et al. 2020) has been proposed. Instead of surface overlap, it measures the pairwise cosine similarities between the contextual token embeddings from the BERT model. T. Zhang et al. (2020) show that BERTScore correlates better with human judgements than metrics based on surface overlap.

These metrics can give an idea of how well the generated explanations reflect the human-written explanations from a dataset. However, although better than previous metrics, even the BERTScore metrics are far from perfect: In particular, even they have been shown to be overly reliant on lexical overlap (Hanna and Bojar 2021).

A more general problem is that the reference explanations are expected to have large variance between annotators, given that the possible set of valid explanations is large. This may induce annotator bias into the evaluation (Geva et al. 2019). Therefore evaluation approaches that compare generations to a gold standard are not considered suitable as a sole base for the evaluation of natural language generation tasks (Amidei et al. 2018).

To judge the plausibility of the generated explanations independent of specific human annotations, it has therefore become common to include a human evaluation of a subset of the evaluation set (Marasovic et al. 2022; Wiegreffe et al. 2022). However, while human ratings are often treated as a gold standard, even they have been shown to rely on surface qualities of the text such as fluency (E. Clark et al. 2021). Moreover, the inter-rater agreement in human evaluations is often low. Amidei et al. (2018) study what influences the agreement and find that ratings diverge for generation tasks due to inter-personal differences in factors such as personal style, attention to detail, background knowledge and personal assumptions. They conclude that such differences are inherent to the nature of such studies, and that the inter-rater agreement is not a measure that is supposed to be maximised.

Faithfulness

As I have outlined in the introduction of this chapter, the faithfulness of a self-rationalising model is, by default, uncertain. There is no inherent reason to assume that the generated explanations accurately reflect the models' prediction process. Therefore, a line of work has emerged that estimates the explanations' faithfulness to the label prediction process.

Hase et al. (2020) propose the *leakage-adjusted simulatability* (LAS) metric to evaluate how predictable (*simulatable*) a model's decisions are to an observer, given the explanations. They argue that the LAS score is a measurement of faithfulness because it is an indication that the content of the explanations influences the predictions. The authors control for predictability from the input alone by comparing to an input-only baseline, as well as for label leakage in the explanation by testing if the label can be predicted with the explanation alone. They find that rationalising models achieve better LAS scores than reasoning model, and that the serial-task rationalising model gets similar scores as humans.

Overall, experimental faithfulness evaluations based on input interventions have lead to mixed conclusions. Wiegreffe et al. (2021) show that there are correlations of the label prediction and the explanation generation process by adding noise to the input and showing that both are affected in similar ways. They conclude that their results indicate the potential to generate faithful free-text explanations. Atanasova et al. (2023) propose analysing the faithfulness of textual explanations after counterfactual interventions on the input that alter the prediction. They find that for many instances, it is possible to find an edit that leads to an unfaithful explanation. They also test for the sufficiency of the reasons provided in the explanations, again finding a substantial number of unfaithful explanations. Turpin et al. (2023) add biasing features to the input which affect the prediction, e.g. providing few-shot multiple-choice examples where the selected answer is always the first one. Those biases are never reflected in the explanations generated by the models they test, showing that the predictions are not faithful to the explanations. Critiquing the aforementioned works, Parcalabescu and Frank (2024) argue that what they measure is not faithfulness but *self-consistency*, which is a necessary, but not a sufficient condition for faithfulness. They argue that the inner workings for generating predictions and explanations could be consistent, but still different, and that therefore, evaluating faithfulness is still an open problem.

An alternative approach to estimating the faithfulness of explanations is comparing the performance of models that include explanations to models that do not provide explanations for their predictions. If self-rationalisation adds performance and robustness, this is an indication that the prediction process is at least partly guided by the reasoning provided in the explanations. Ross et al. (2022) test the effect of self-rationalisation on the robustness to spurious correlations in fine-tuned models by evaluating on challenging subsets and datasets that are designed for not containing known spurious correlations in NLI (Gururangan et al. 2018; T. McCoy et al. 2019). They find that it does not generally make models more robust. However, they identify two factors that make a model more likely to benefit from self-rationalisation: fewer fine-tuning resources and larger model sizes.

4.3.4 Summary

Natural language explanations can cover a broad set of problems and are commonly more accessible to humans than other forms of explanation. As the capabilities of LLMs have increased and few-shot and zero-shot prompting has become possible, self-rationalisation is now a simple and powerful method to generate natural language explanations. However, the properties of free-text explanations are still underexplored. Our work in Papers IV and V aims at understanding their role and effects better.

5 Paper Summaries

In this chapter, I will summarise the contributions of the five papers included in this thesis in Sections 5.1 to 5.5. The published versions of the full papers are included in Part II; they are adjusted in formatting but the text is not altered. In Section 5.6, I will give an overview of other papers that I have contributed to during my time as a PhD student. These papers are not an official part of this thesis. Although the main contributions of two of these papers are outside this field, I will specifically reflect on their relation to interpretable and explainable NLP.

5.1 Paper I

Jenny Kunz and Marco Kuhlmann (Dec. 2020). "Classifier Probes May Just Learn from Linear Context Features". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5136–5146. DOI: 10.18653/v1/2020.coling-main.450. URL: https://aclanthology.org/2020.coling-main.450.

In this paper, we question an assumption that was often stated in early probing work: That the contextual representation of one token does not contain sufficient information for a probe to learn common sentence-level probing tasks without specific linguistic structure already being *encoded* in the representation. However, through the contextualisation with the multi-head self-attention mechanism, the representation can contain detailed information about the whole sentence. This information enables the learning of all sentence-level tasks as crucial features (in particular, information about the surrounding words) are present.

We show experimentally that detailed linear context features are actually contained in the representation by constructing a probing task that, from the representation of one token at a time, predicts the *exact* identity of the neighbouring words at a specific position. Our results show that for the BERT model, more than half of the direct neighbours of the word in the sentence are recoverable exactly, and even more distant neighbours are recoverable to a substantial extent. For ELMo, the numbers are lower but still much higher than for non-contextual baselines.

Based on these results, we develop a framework that is centred around what we call the *context-only hypothesis*: We call for probing experiments to disprove the assumption that the only information that the probe uses to learn the probing task is linear information about the identity of the neighbouring words. We show theoretically and experimentally that none of the baselines or restrictions imposed on the classifier in the literature can disprove this hypothesis. Therefore, we argue that these methods are not a strong foundation for conclusions about linguistic features being explicitly encoded in the representations.

5.2 Paper II

Jenny Kunz and Marco Kuhlmann (Nov. 2021). "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, pp. 15–25. DOI: 10.18653/v1/2021.blackboxnlp-1.2. URL: https://aclanthology.org/2021.blackboxnlp-1.2.

In this paper, we take inspiration from our first paper and increase the restrictiveness of the probing classifier. As machine learning models typically operate within an *interpolation* setting, we hypothesise that the ability to *extrapolate* would be an indication that the probe *decodes* linguistic knowledge rather than learning the task from scratch. For this purpose, we take inspiration from curriculum learning and define a set of linguistic, statistical and learning-related scoring functions after which we rank the difficulty of training examples. These rankings are used to define train-test splits where we train only on easy examples and evaluate only on hard examples.

We analyse the relative merits of these criteria experimentally and theoretically, finding that the results vary greatly across scoring functions. The linguistic and statistical criteria that allow for the best-motivated splitting points show the clearest differences between interpolation and extrapolation settings, indicating that arbitrary splitting points do not define true extrapolation setups. The most discriminating ranking function for the syntactic dependency label prediction probing task is the length of the syntactic dependency arc, where the easy samples are the samples where the syntactic head is a direct neighbour. For the part-of-speech-tagging probing task, the most differentiating criterion is a binary statistical criterion, where the easy samples are the ones where a word form is assigned its most frequent label. Our experiments show that for the setups with well-motivated splitting points, the ability to extrapolate to harder examples is limited, but not completely absent.

5.3 Paper III

Jenny Kunz and Marco Kuhlmann (Oct. 2022). "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, pp. 4664–4676. URL: https://aclanthology.org/2022.coling-1.413.

The third paper also proposes a way to overcome the question if a task is encoded in the representation or learned by the probe. We offer a new perspective on how to approach, evaluate and interpret probing results with a new family of metrics that focuses on the information flow through the model instead of the isolated performance of one layer at a time. The new metrics focus on the local *information gain* from one layer to the other, and calculate the *contribution* of this layer to the overall performance of the model.

This new focus shows that it is the very first layers of the BERT model that contribute the most new information to the performance on syntactic tasks, and not the middle layers as one may conclude based on traditional, global probing metrics. The information is just preserved in the model, which leads to a higher performance due to more accumulated information.

We also test if expected hierarchies between probing tasks hold: That information useful for predicting syntactic parents is gained in earlier layers than information relevant for syntactic grandparent prediction, and that information for predicting the most frequent part-of-speech tags for a word form is gained earlier than information for less frequent tags, as those may require more context. Our experimental results show that the hierarchy holds in the new metrics only for the first task pair; for the less frequent part-of-speech tags, the most significant gains happen in the very first layers, earlier than for the most frequent tags. We also show that the structure of information varies greatly between models in different languages.

5.4 Paper IV

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann (Dec. 2022). "Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions". In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 164–177. DOI: 10.18653/v1/2022.blackboxnlp-1.14. URL: https://aclanthology.org/2022.blackboxnlp-1.14. **Best Paper Award Winner.**

The fourth paper is our first contribution to the field of natural language explanations. We explore how the utility of explanations for a downstream model compares to human ratings of the same explanations by training and evaluating explanation-generating models with two different pipeline architectures: one serial-task model that generates only the explanation in the first step, and one multi-task model that jointly

generates predictions with the explanations. Both pipelines consist of a fine-tuned GPT-2 for the generation of the explanations and BERT for the final classification, and are trained and evaluated on e-SNLI and ECQA.

We find that the explanations by the serial-task model are rated higher by humans with respect to their validity and factual correctness and score higher in similarity measures compared to human-annotated gold explanations. However, in the utility for a downstream classification model, the explanations by the multi-task model score slightly higher. We observe that the latter explanations are a classical example of hallucinations, as they contain more novel information and more factually incorrect information. This is punished by human annotators but slight performance improvements may be an indicator that the novel information can in some cases be useful context for downstream models.

A second insight from this paper is that fine-tuning the downstream classifier with gold explanations only leads to failure at evaluation time when switching to generated explanation. This indicates an over-reliance on information in the explanations. However, further fine-tuning on generated explanations solves this problem. The classifier learns to handle the less reliable information from the generated explanations and, in the case of the harder ECQA dataset, improves its performance over the model that does not receive explanations as an additional input substantially.

5.5 Paper V

Jenny Kunz and Marco Kuhlmann (2024). *Properties and Challenges of LLM-Generated Explanations*. arXiv: 2402.10532 [cs.CL]. URL: https://arxiv.org/abs/2402.10532. Under Review.

In the fifth paper, we focus on zero-shot explanations generated by GPT-4 and explore their properties. We annotate an instruction fine-tuning dataset for categories of instruction, and the outputs of GPT-4 for known properties of human explanations that have been pointed out as problematic for the goals of explainable NLP.

We find that the explanations often list an incomplete set of contributing reasons, but argue that this is not avoidable due to the open and complex nature of many instructions. We also find a large number of explanations that contain illustrative elements, which likely have no connection to the model's reasoning but on the other hand help make the reasoning more accessible to the user. Only few explanations are subjective, which we assume is related to GPT-4's alignment process that discourages the inclusion of subjective statements in the output. Misleading explanations for answers that are wrong are also rare, but we assume that this is an artefact of the dataset we use, where the instructions themselves are LLM-generated and therefore very likely to be answerable correctly by GPT-4.

We discuss the effects of the presence or absence of these properties on different goals of explainable NLP and different user groups that may use LLMs. We argue that all properties can have positive or negative sides, depending on the use case.

5.6 Other Works

During my PhD studies, I also contributed to the following papers.

5.6.1 Constructing Surrogate Models for Textual Explanations

Marc Braun and Jenny Kunz (2024). *A Hypothesis-Driven Framework for the Analysis of Self-Rationalising Models*. arXiv: 2402.04787 [cs.CL]. URL: https://arxiv.org/abs/2402.04787 To Appear at the Student Research Workshop at the 18th Conference of the European Chapter of the Association for Computational Linguistics.

In this work, we build surrogate models based on a Bayesian Network that allow us to test hypotheses on how a task (in our case, NLI) is solved. The Bayesian Network models interactions between phrases of the premise and hypothesis in an interpretable way, where its internal states can be translated to textual explanations via templates.

We compare the explanations generated with the surrogate models to the explanations generated by few-shot-prompted GPT-3.5. We find that the explanations can take similar shapes to the ones in the e-SNLI dataset, but that the performance of the Bayesian Network model is too low to reach a high similarity to GPT-3.5 in quantitative metrics. Due to the template-like structure of e-SNLI explanations, we do however see the potential to better approximate the GPT-3.5 explanations and decisions with a more sophisticated surrogate model.

5.6.2 Understanding Cross-Lingual Transfer

Another line of work analyses how cross-lingual transfer is done in LLMs. Two papers have so far come out of this side project. The first one investigates if GPT-3.5's generations are language-specific or if the same capabilities are shared across languages. In the second one, we perform extensive ablations on the contribution of language adapters to zero-shot cross-lingual transfer.

Transfer of Capabilities across Languages: Frameworks and Evaluation

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann (May 2023). "Bridging the Resource Gap: Exploring the Efficacy of English and Multilingual LLMs for Swedish". In: *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*. Tórshavn, the Faroe Islands: Association for Computational Linguistics, pp. 92–110. URL: https://aclanthology.org/2023.resourceful-1.13.

In our first paper on cross-lingual transfer, we compare GPT-3.5's English and Swedish abilities concerning whether capabilities are shared across languages, or if distinct submodels are activated depending on the input language.

We find that many capabilities are shared: With Swedish input, the performance is relatively close to GPT-3.5 in English, and differences can in part be explained by translation errors from the evaluation set. As other LLMs with substantially more Swedish

pre-training data have much more limited capabilities on the same tasks, we assume that the Swedish submodel would either be too small or not have the appropriate training data for these capabilities to be learned. Therefore we conclude that those capabilities must be shared with or transferred to the lower-resource languages in the model.

However, we still find a distinct behaviour of GPT-3.5 when prompted in Swedish as compared to English. A small study with prompts that we hypothesise can trigger culture-specific answers shows that indeed, the output reflects common cultural practices of Sweden versus the United States of America. We conclude that the transfer behaviour that GPT-3.5 shows is promising; with transfer happening where it is desired but with specific outputs where appropriate.

Language Adapters as Modular Components

Jenny Kunz and Oskar Holmström (2024). *The Impact of Language Adapters in Cross-Lingual Transfer for NLU*. arXiv: 2402.00149 [cs.CL]. URL: https://arxiv.org/abs/2402.00149 To Appear at the First Workshop on Modular and Open Multilingual NLP (MOOMIN) at the 18th Conference of the European Chapter of the Association for Computational Linguistics.

In the second paper on cross-lingual transfer, we explore the effect of language adapters when performing zero-shot transfer for natural language understanding tasks. Specifically, we train task adapters with the language adapter of the source language active and exchange the source-language adapter against the target-language adapter at evaluation time.

We perform an extensive set of ablations on the impact of the target-language adapters. We observe that it is often possible to keep the source-language adapter at evaluation time instead or even to remove the language adapter without substitution without a strong negative effect. Sometimes, keeping the source-language adapter even outperforms using the target-language adapter. For the context of interpretable NLP, this suggests that language adapters trained independently of the model do not play a consistent, and thereby interpretable, role.

6 Conclusion

In this final chapter, I will summarise the contributions of this thesis and set them into a wider context in Section 6.1. Finally, in Section 6.2, I will provide an outlook concerning current trends and developments in the field of interpretable and explainable NLP.

6.1 Summary

In this thesis, I have presented our contributions to enhancing interpretability techniques and better understanding the properties of natural language explanations generated by LLMs.

Our contributions to the field of interpretability focused on probing classifiers that measure linguistic information that the internal representation of the model, often from a specific layer, contains. We developed a framework for a rigid assessment of the results of probing classifiers. A crucial distinction that we discuss in this paper is if the probing classifier *decodes* information from the representation or if it *learns* the probing task. We show that a classical argument why probes should not be able to do the latter does not hold and that common restrictions to the probing methodology are not able to make the distinction either. Building on these results, we developed more challenging probing methods that focus on the ability of the probe to extrapolate from easy to hard examples. We argue that as classifiers generally operate in an interpolation setup, extrapolation capabilities would be an indication that the probe has found generalisable features from the representation. We designed new evaluation metrics that focus on modelling local information gains throughout the model and each layer's contribution to the model's overall performance. This shifts the focus on

where syntactic information is located in the model from the middle layers, where the overall performance is the highest, to the earliest layers, which contribute the most new information. Both the focus on extrapolation capabilities and the focus on local information gains provide new perspectives to the probing method which can help us to gain more robust insights.

In the field of explainability, we focused on the properties and utility of free-text explanations generated by LLMs. We showed that human ratings, particularly of the factual correctness of the explanations, are not indicative of the performance when using the generated explanations in a downstream model. For the latter, including more novel information in the explanations appears to be beneficial, but this comes at the risk of including incorrect or irrelevant information, which human raters punish decisively. We further examined which common properties of human explanations are commonly reflected in LLM-generated explanations. The results of our annotation study indicate that they often list incomplete sets of contributing reasons as well as illustrative examples.

Interpretation and explanation methods can complement each other in shaping a better understanding of LLMs. The former methods give us a high-level understanding of how the individual tokens are contextualised and, layer for layer, form a representation useful for many applications. The latter methods give us an idea of the context and reasoning accessible to the model when making a prediction, even if the explanations are not faithful to the model's decision process. Together with an understanding of the LLMs' architecture and training objectives, such methods make it possible to achieve a coarse understanding of the decision-making process and be able to predict the models' behaviour to a certain extent. This understanding is insufficient to allow for the deployment of LLMs for high-stakes decisions without a human in the loop. However, it has the potential to enable developers and users to make better decisions on whether the model will be able to perform a certain task, whether its decision-making is sufficiently robust, and how it can be improved.

6.2 Outlook

Finally, I outline some questions that I consider central for future work in interpretable and explainable NLP, based on our research on these topics.

6.2.1 Understanding the Internal Processes of LLMs

Probing classifiers have given us valuable insights into how linguistic information is structured within an LLM and how this representation is formed during the training of the model. We have made contributions to this method in Papers I, II and III. However, like with many other interpretability methods, as outlined in Chapter 3, those insights are coarse and need to be interpreted relative to a baseline or to other models or layers. An open question is if it is possible to reach a more general and fine-grained understanding that is still comprehensible to human observers.

Recently, *mechanistic interpretability* has attracted the attention of many researchers, with some successes reported on the reverse engineering of toy models. Large amounts

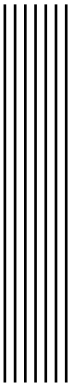
of resources are flowing into this field from both industry and academia. As I discussed in Chapter 3.3, results in NLP are so far focused on specific features rather than algorithm discovery and are far from the declared goal of meaningfully explaining decision-making in any real-world application. Whether even this field hits a wall with toy tasks and standalone features or whether it can give us more fine-grained mechanistic insights into how LLMs work remains to be seen.

6.2.2 Building LLMs that are More Interpretable by Design

Another promising line of research is building coarsely interpretable models without losing the capabilities LLMs are appreciated for. Such models could be trained to have modules that fulfill specific interpretable functions. Pfeiffer et al. (2022) have successfully employed such a modular approach for language modules in encoder models. Unlike our own experiments with language adapters trained post-hoc, as summarised in Section 5.6.2, such modules that are present already at pre-training time may be an isolated encapsulation of a specific property. It would be insightful, although more complex, to test a similar approach for capabilities other than handling different input languages. Such models are not fully explainable in that all details of the decision-making process are comprehensible but offer many insights that are not possible with LLMs by being more controllable as they allow for targeted interventions.

6.2.3 Targeted Explanations that Consider the User's Needs

As we have outlined in Papers IV and V, explanations can have different goals and target groups, and properties beneficial for one user group can be disadvantageous for another. However, research that explicitly states the intended target group for the explanations, let alone specifies their needs, is sparse in explainable NLP research. This ignorance contributes to the fact that currently, end users of the systems rarely engage with explanations and either ignore the system's recommendations entirely or blindly follow its predictions (Miller 2023). It has become clear that it is not sufficient for researchers and developers to follow their intuitions on how to design explainable systems (maybe unless the target group is themselves). Therefore, two challenges need to be solved: First, the technical challenge to make generated explanations reflect the decision-making process to an extent sufficient for the problem. And second, the challenge to design the explainable system in a way that reflects the actual target group's needs and is accessible to them.



Bibliography

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg (2017). "Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=BJh6Ztuxl.

Aggarwal, Shourya, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg (Aug. 2021). "Explanations for CommonsenseQA: New Dataset and 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, pp. 3050–3065. DOI: 10.18653/v1/2021.acl-long.238. URL: https://aclanthology.org/2021.acl-long.238.

Alain, Guillaume and Yoshua Bengio (2017). "Understanding intermediate layers using linear classifier probes". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=HJ4-rAVtl.

Alvarez-Melis, David and Tommi S. Jaakkola (2018). "On the Robustness of Interpretability Methods". In: cite arxiv:1806.08049Comment: presented at 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden. URL: http://arxiv.org/abs/1806.08049.

Amidei, Jacopo, Paul Piwek, and Alistair Willis (Aug. 2018). "Rethinking the Agreement in Human Evaluation Tasks". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 3318–3329. URL: https://aclanthology.org/C18-1281.

Ananthakrishnan, R, Pushpak Bhattacharyya, M Sasikumar, and Ritesh M Shah (2007). "Some issues in automatic evaluation of English-Hindi MT: More blues for BLEU". In: *Icon* 64.

Atanasova, Pepa, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein (July 2023). "Faithfulness Tests for Natural Language Explanations". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 283–294. DOI: 10.18653/v1/2023.acl-short.25. URL: https://aclanthology.org/2023.acl-short.25.

Atanasova, Pepa, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein (Nov. 2020a). "A Diagnostic Study of Explainability Techniques for Text Classification". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3256–3274. DOI: 10.18653/v1/2020.emnlp-main.263. URL: https://aclanthology.org/2020.emnlp-main.263.

— (July 2020b). "Generating Fact Checking Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7352–7364. DOI: 10.18653/v1/2020.acl-main.656. URL: https://aclanthology.org/2020.acl-main.656.

Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni (2023). *Factuality Challenges in the Era of Large Language Models*. arXiv: 2310.05189 [cs.CL].

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1409.0473.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld (2021). "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing

Machinery. ISBN: 9781450380966. DOI: 10.1145/3411764.3445717. URL: https://doi.org/10.1145/3411764.3445717.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (June 2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 238–247. DOI: 10.3115/v1/P14-1023. URL: https://aclanthology.org/P14-1023.

Bastings, Jasmijn and Katja Filippova (Nov. 2020). "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?" In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, pp. 149–155. DOI: 10.18653/v1/2020.blackboxnlp-1.14. URL: https://aclanthology.org/2020.blackboxnlp-1.14.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass (July 2017a). "What do Neural Machine Translation Models Learn about Morphology?" In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 861–872. DOI: 10.18653/v1/P17-1080. URL: https://aclanthology.org/P17-1080.

Belinkov, Yonatan and James Glass (2019). "Analysis Methods in Neural Language Processing: A Survey". In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. DOI: 10.1162/tacl_a_00254. URL: https://aclanthology.org/Q19-1004.

Belinkov, Yonatan, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass (Nov. 2017b). "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, pp. 1–10. URL: https://aclanthology.org/I17-1001.

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

Blevins, Terra, Hila Gonen, and Luke Zettlemoyer (Dec. 2022). "Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3575–3590. DOI: 10.18653/v1/2022.emnlp-main.234. URL: https://aclanthology.org/2022.emnlp-main.234.

Blevins, Terra, Omer Levy, and Luke Zettlemoyer (July 2018). "Deep RNNs Encode Soft Hierarchical Syntax". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 14–19. DOI: 10.18653/v1/P18-2003. URL: https://aclanthology.org/P18-2003.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. (2021). "On the Opportunities and Risks of Foundation Models". In: *arXiv preprint arXiv:2108.07258*.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (Sept. 2015). "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, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://aclanthology.org/D15-1075.

Braun, Marc and Jenny Kunz (2024). *A Hypothesis-Driven Framework for the Analysis of Self-Rationalising Models*. arXiv: 2402.04787 [cs.CL]. URL: https://arxiv.org/abs/2402.04787.

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan,

and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom (2018). "e-SNLI: Natural Language Inference with Natural Language Explanations". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf.

Cammarata, Nick, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim (2020). "Thread: Circuits". In: *Distill*. https://distill.pub/2020/circuits. DOI: 10.23915/distill.00024.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson (2013). "One billion word benchmark for measuring progress in statistical language modeling". In: *arXiv preprint arXiv:1312.3005*.

Chen, Howard, Jacqueline He, Karthik Narasimhan, and Danqi Chen (July 2022). "Can Rationalization Improve Robustness?" In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 3792–3805. DOI: 10.18653/v1/2022.naacl-main.278. URL: https://aclanthology.org/2022.naacl-main.278.

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

Chu, Eric, Deb Roy, and Jacob Andreas (2020). "Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction". In: *CoRR*

abs/2007.12248. arXiv: 2007.12248. URL: https://arxiv.org/abs/2007.12248.

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith (Aug. 2021). "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 7282–7296. DOI: 10.18653/v1/2021.acl-long.565. URL: https://aclanthology.org/2021.acl-long.565.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning (Aug. 2019). "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://aclanthology.org/W19-4828.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman (2021). "Training Verifiers to Solve Math Word Problems". In: *CoRR* abs/2110.14168. arXiv: 2110.14168. URL: https://arxiv.org/abs/2110.14168.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loı¨c Barrault, and Marco Baroni (July 2018). "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, pp. 2126–2136. DOI: 10.18653/v1/P18-1198. URL: https://aclanthology.org/P18-1198.

Cummins, Robert C. (2000). ""How Does It Work" Versus "What Are the Laws?": Two Conceptions of Psychological Explanation". In: *Explanation and Cognition, 117-145*. Ed. by F. Keil and Robert A. Wilson. MIT Press. URL: https://philpapers.org/rec/CUMHDI.

Dalvi, Fahim, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass (2019). "What is one grain of sand in the desert? analyzing individual neurons in deep NLP 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*. AAAI'19/IAAI'19/EAAI'19. Honolulu, Hawaii, USA: AAAI Press. ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33016309. URL: https://doi.org/10.1609/aaai.v33i01.33016309.

DeJong, Gerald and Raymond Mooney (1986). "Explanation-based learning: An alternative view". In: *Machine learning* 1.2, pp. 145–176. URL: https://link.springer.com/article/10.1007/BF00114116.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace (July 2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (Nov. 2021). "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1286–1305. DOI: 10.18653/v1/2021.emnlp-main.98. URL: https://aclanthology.org/2021.emnlp-main.98.

Dombrowski, Ann-Kathrin, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel (2019). "Explanations can be manipulated and geometry is to blame." In: *NeurIPS*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, pp. 13567–13578. URL: http://dblp.uni-trier.de/db/conf/nips/nips2019.html#DombrowskiAAAMK19.

Dozat, Timothy, Peng Qi, and Christopher D. Manning (Aug. 2017). "Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 20–30. DOI: 10.18653/v1/K17-3002. URL: https://aclanthology.org/K17-3002.

Dziri, Nouha, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose (Nov. 2021). "Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2197–2214. DOI: 10.18653/v1/2021.emnlp-main.168. URL: https://aclanthology.org/2021.emnlp-main.168.

Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg (2021). "Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals". In: *Transactions of the Association for Computational Linguistics* 9, pp. 160–175. DOI: 10.1162/tacl_a_00359. URL: https://aclanthology.org/2021.tacl-1.10.

Elman, Jeffrey L (1990). "Finding structure in time". In: *Cognitive science* 14.2, pp. 179–211.

Emerson, Guy (July 2020). "What are the Goals of Distributional Semantics?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7436–7453. DOI: 10.18653/v1/2020.acl-main.663. URL: https://aclanthology.org/2020.acl-main.663.

Ethayarajh, Kawin (Nov. 2019). "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 55–65. DOI: 10.18653/v1/D19-1006. URL: https://aclanthology.org/D19-1006.

Firth, J. R. (1957). "A synopsis of linguistic theory 1930-55." In: 1952-59, pp. 1–32.

Geva, Mor, Yoav Goldberg, and Jonathan Berant (Nov. 2019). "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1161–1166. DOI: 10.18653/v1/D19-1107. URL: https://aclanthology.org/D19-1107.

Gokaslan, Aaron, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex (2019). *OpenWebText Corpus*. URL: http://Skylion007.github.io/OpenWebTextCorpus.

Goldberg, Yoav (2019). "Assessing BERT's syntactic abilities". In: *arXiv preprint arXiv:1901.05287*.

González, Ana Valeria, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer (Aug. 2021). "Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 1103–1116. DOI: 10.18653/v1/2021.findings-acl.95. URL: https://aclanthology.org/2021.findings-acl.95.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (June 2018). "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, pp. 107–112. DOI: 10.18653/v1/N18-2017. URL: https://aclanthology.org/N18-2017.

Hanna, Michael and Ondřej Bojar (Nov. 2021). "A Fine-Grained Analysis of BERTScore". In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 507–517. URL: https://aclanthology.org/2021.wmt-1.59.

Harnad, Stevan (1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1-3, pp. 335–346.

Harris, Zellig S. (1954). "Distributional Structure". In: WORD 10.2-3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

Hase, Peter, Shiyue Zhang, Harry Xie, and Mohit Bansal (Nov. 2020). "Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?" In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4351–4367. DOI: 10.18653/v1/2020.findings-emnlp.390. URL: https://aclanthology.org/2020.findings-emnlp.390.

He, Xuanli, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp (2023). *Using Natural Language Explanations to Improve Robustness of In-context Learning for Natural Language Inference*. arXiv: 2311.07556 [cs.CL].

Hewitt, John and Percy Liang (Nov. 2019). "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, pp. 2733–2743. DOI: 10.18653/v1/D19-1275. URL: https://aclanthology.org/D19-1275.

Hewitt, John and Christopher D. Manning (June 2019). "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, pp. 4129–4138. DOI: 10.18653/v1/N19-1419. URL: https://aclanthology.org/N19-1419.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Holmström, Oskar, Jenny Kunz, and Marco Kuhlmann (May 2023). "Bridging the Resource Gap: Exploring the Efficacy of English and Multilingual LLMs for Swedish". In: *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*. Tórshavn, the Faroe Islands: Association for Com-

putational Linguistics, pp. 92–110. URL: https://aclanthology.org/2023.resourceful-1.13.

Hong, Ruixin, Hongming Zhang, Xintong Yu, and Changshui Zhang (July 2022). "METGEN: A Module-Based Entailment Tree Generation Framework for Answer Explanation". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, pp. 1887–1905. DOI: 10.18653/v1/2022.findings-naacl.145. URL: https://aclanthology.org/2022.findings-naacl.145.

Howard, Jeremy and Sebastian Ruder (July 2018). "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://aclanthology.org/P18-1031.

Hsieh, Yu-Lun, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh (July 2019). "On the Robustness of Self-Attentive Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1520–1529. DOI: 10.18653/v1/P19-1147. URL: https://aclanthology.org/P19-1147.

Hupkes, Dieuwke, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. (2022). "State-of-the-art generalisation research in NLP: a taxonomy and review". In: *arXiv preprint arXiv:2210.03050*.

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema (2017). "Visualisation and 'Diagnostic Classifiers' Reveal how Recurrent and Recursive Neural Networks Process Hierarchical Structure." In: *Interpreting, Explaining and Visualizing Deep Learning, NIPS2017*.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry (2019). "Adversarial Examples Are Not Bugs, They Are Features". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

Jacovi, Alon and Yoav Goldberg (July 2020). "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, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386.

Jain, Sarthak and Byron C. Wallace (June 2019). "Attention is not Explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the As-*

sociation for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://aclanthology.org/N19-1357.

Jawahar, Ganesh, Benoit Sagot, and Djame Seddah (July 2019). "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluis Marquez. Florence, Italy: Association for Computational Linguistics, pp. 3651–3657. DOI: 10.18653/v1/P19-1356. URL: https://aclanthology.org/P19-1356.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung (2023). "Survey of Hallucination in Natural Language Generation". In: *ACM Comput. Surv.* 55.12. ISSN: 0360-0300. DOI: 10.1145/3571730. URL: https://doi.org/10.1145/3571730.

Jie, Zhanming, Jierui Li, and Wei Lu (May 2022). "Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5944–5955. DOI: 10.18653/v1/2022.acl-long.410. URL: https://aclanthology.org/2022.acl-long.410.

Karpathy, Andrej, Justin Johnson, and Li Fei-Fei (2015). "Visualizing and Understanding Recurrent Networks". In: *CoRR* abs/1506.02078. arXiv: 1506.02078. URL: http://arxiv.org/abs/1506.02078.

Keil, Frank C (2006). "Explanation and understanding". In: *Annual review of psychology* 57, p. 227. DOI: 10.1146/annurev.psych.57.102904.190100. URL: https://www.annualreviews.org/doi/10.1146/annurev.psych.57.102904.190100.

Kiperwasser, Eliyahu and Yoav Goldberg (2016). "Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations". In: *Transactions of the Association for Computational Linguistics* 4, pp. 313–327. DOI: 10.1162/tacl_a_00101. URL: https://aclanthology.org/Q16-1023.

Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko (2023). *ChatGPT: Jack of all trades, master of none*. DOI: 10.48550/ARXIV.2302.10724. URL: https://arxiv.org/abs/2302.10724.

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large Language Models are Zero-Shot Rea-

soners". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton (2019). "Similarity of Neural Network Representations Revisited". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3519–3529. URL: https://proceedings.mlr.press/v97/kornblith19a.html.

Kunz, Jenny and Oskar Holmström (2024). *The Impact of Language Adapters in Cross-Lingual Transfer for NLU*. arXiv: 2402.00149 [cs.CL]. URL: https://arxiv.org/abs/2402.00149.

Kunz, Jenny, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann (Dec. 2022). "Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions". In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 164–177. DOI: 10.18653/v1/2022.blackboxnlp-1.14. URL: https://aclanthology.org/2022.blackboxnlp-1.14.

Kunz, Jenny and Marco Kuhlmann (Dec. 2020). "Classifier Probes May Just Learn from Linear Context Features". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5136–5146. DOI: 10.18653/v1/2020.coling-main.450. URL: https://aclanthology.org/2020.coling-main.450.

— (Nov. 2021). "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, pp. 15–25. DOI: 10.18653/v1/2021.blackboxnlp-1.2. URL: https://aclanthology.org/2021.blackboxnlp-1.2.

— (Oct. 2022). "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, pp. 4664–4676. URL: https://aclanthology.org/2022.coling-1.413.

— (2024). *Properties and Challenges of LLM-Generated Explanations*. arXiv: 2402.10532 [cs.CL]. URL: https://arxiv.org/abs/2402.10532.

Lakatos, Imre (1963). *Proofs and refutations*. Nelson London.

Latcinnik, Veronica and Jonathan Berant (2020). "Explaining Question Answering Models through Text Generation". In: *CoRR* abs/2004.05569. arXiv: 2004.05569. URL: https://arxiv.org/abs/2004.05569.

Lee, Eunjin, David Braines, Mitchell Stiffler, Adam Hudler, and Daniel Harborne (2019). "Developing the sensitivity of LIME for better machine learning explanation". In: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*. Vol. 11006. SPIE, pp. 349–356. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/1100610/Developing-the-sensitivity-of-LIME-for-better-machine-learning-explanation/10.1117/12.2520149.full.

Lee, Katherine, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo (2018). "Hallucinations in neural machine translation". In: *Interpretability and Robustness in Audio, Speech, and Language Workshop. Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.* URL: https://openreview.net/pdf?id=SJxTk3vB3m.

Levy, Omer and Yoav Goldberg (June 2014). "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 171–180. DOI: 10.3115/v1/W14-1618. URL: https://aclanthology.org/W14-1618.

Lewis, Clayton (1988). "Why and How to Learn Why: Analysis-based Generalization of Procedures". In: *Cognitive Science* 12.2, pp. 211–256. DOI: https://doi.org/10.1207/s15516709cog1202_3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1202_3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1202_3.

Lhoneux, Miryam de, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre (Aug. 2017). "From Raw Text to Universal Dependencies - Look, No Tags!" In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 207–217. DOI: 10.18653/v1/K17-3022. URL: https://aclanthology.org/K17-3022.

Li, Jiwei, Xinlei Chen, Eduard Hovy, and Dan Jurafsky (June 2016). "Visualizing 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, pp. 681–691. DOI: 10.18653/v1/N16-1082. URL: https://aclanthology.org/N16-1082.

Li, Yixuan, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft (2015). "Convergent Learning: Do different neural networks learn the same representations?" In: *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*. Ed. by Dmitry

Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar. Vol. 44. Proceedings of Machine Learning Research. Montreal, Canada: PMLR, pp. 196–212. URL: https://proceedings.mlr.press/v44/li15convergent.html.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://aclanthology.org/W04-1013.

Lin, Yongjie, Yi Chern Tan, and Robert Frank (Aug. 2019). "Open Sesame: Getting inside BERT's Linguistic Knowledge". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 241–253. DOI: 10.18653/v1/W19-4825. URL: https://aclanthology.org/W19-4825.

Ling, Wang, Dani Yogatama, Chris Dyer, and Phil Blunsom (July 2017). "Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 158–167. DOI: 10.18653/v1/P17-1015. URL: https://aclanthology.org/P17-1015.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016). "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies". In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535. DOI: 10.1162/tacl_a_00115. URL: https://aclanthology.org/Q16-1037.

Lipton, Zachary C (2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.

Lombrozo, Tania (2006). "The structure and function of explanations". In: *Trends in cognitive sciences* 10.10, pp. 464–470. URL: https://pubmed.ncbi.nlm.nih.gov/16942895/.

Luccioni, Alexandra Sasha and Anna Rogers (2023). *Mind your Language (Model): Fact-Checking LLMs and their Role in NLP Research and Practice*. arXiv: 2308.07120 [cs.CL].

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964.

Magar, Inbal and Roy Schwartz (May 2022). "Data Contamination: From Memorization to Exploitation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 157–165. DOI: 10.18653/v1/2022.acl-short.18. URL: https://aclanthology.org/2022.acl-short.18.

Malmström, Hans, Christian Stöhr, and AW Ou (2023). "Chatbots and other AI for learning: A survey of use and views among university students in Sweden". In: *Chalmers Studies in Communication and Learning in Higher Education* 1.10.17196. URL: https://research.chalmers.se/publication/535715/file/535715_Fulltext.pdf.

Marasovic, Ana, Iz Beltagy, Doug Downey, and Matthew Peters (July 2022). "Few-Shot Self-Rationalization with Natural Language Prompts". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, pp. 410–424. DOI: 10.18653/v1/2022.findings-naacl.31. URL: https://aclanthology.org/2022.findings-naacl.31.

Marvin, Rebecca and Tal Linzen (Oct. 2018). "Targeted Syntactic Evaluation of Language Models". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1192–1202. DOI: 10.18653/v1/D18-1151. URL: https://aclanthology.org/D18-1151.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173.

McCoy, R Thomas, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths (2023). "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve". In: *arXiv preprint arXiv:2309.13638*. URL: https://arxiv.org/pdf/2309.13638.pdf.

McCoy, Tom, Ellie Pavlick, and Tal Linzen (July 2019). "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, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: https://aclanthology.org/P19-1334.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov (2022). "Locating and Editing Factual Associations in GPT". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 17359–17372. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Efficient estimation of word representations in vector space". In.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013b). "Distributed Representations of Words and Phrases and

their Compositionality". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267, pp. 1–38.

— (2023). "Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Chicago, IL, USA: Association for Computing Machinery, pp. 333–342. DOI: 10.1145/3593013.3594001. URL: https://doi.org/10.1145/3593013.3594001.

Mitchell, Tom M, Richard M Keller, and Smadar T Kedar-Cabelli (1986). "Explanation-based generalization: A unifying view". In: *Machine learning* 1.1, pp. 47–80. URL: https://link.springer.com/article/10.1023/A:1022691120807.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter (2019). "Explaining Explanations in AI". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 279–288. ISBN: 9781450361255. DOI: 10.1145/3287560.3287574. URL: https://doi.org/10.1145/3287560.3287574.

Mollo, Dimitri Coelho and Raphaël Millière (2023). "The vector grounding problem". In: *arXiv preprint arXiv:2304.01481*. URL: https://arxiv.org/pdf/2304.01481.pdf.

Moravcsik, Julius ME (1974). "Aristotle on adequate explanations". In: *Synthese*, pp. 3–17. URL: https://www.jstor.org/stable/pdf/20114949.pdf.

Nanda, Neel, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt (2023). "Progress measures for grokking via mechanistic interpretability". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. URL: https://openreview.net/pdf?id=9XFSbDPmdW.

Narang, Sharan, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan (2020). *WT5?! Training Text-to-Text Models to Explain their Predictions*. arXiv: 2004.14546 [cs.CL].

Narayanan, Menaka, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez (2018). "How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation". In: *CoRR* abs/1802.00682. arXiv: 1802.00682. URL: http://arxiv.org/abs/1802.00682.

Nie, Feng, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin (July 2019). "A Simple Recipe towards Reducing Hallucination in Neural Surface Realisation". In: *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics. Florence, Italy: Association for Computational Linguistics, pp. 2673–2679. DOI: 10.18653/v1/P19-1256. URL: https://aclanthology.org/P19-1256.

Niven, Timothy and Hung-Yu Kao (July 2019). "Probing Neural Network Comprehension of Natural Language Arguments". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4658–4664. DOI: 10.18653/v1/P19-1459. URL: https://aclanthology.org/P19-1459.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter (2020). "Zoom In: An Introduction to Circuits". In: *Distill*. https://distill.pub/2020/circuits/zoom-in. DOI: 10.23915/distill.00024.001.

OpenAI (2023). "GPT-4 Technical Report". In: *ArXiv* abs/2303.08774. URL: https://arxiv.org/abs/2303.08774.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Papenmeier, Andrea, Dagmar Kern, Gwenn Englebienne, and Christin Seifert (2022). "It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI". In: *ACM Trans. Comput.-Hum. Interact.* 29.4. ISSN: 1073-0516. DOI: 10.1145/3495013. URL: https://doi.org/10.1145/3495013.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

Parcalabescu, Letitia and Anette Frank (2024). *On Measuring Faithfulness or Self-consistency of Natural Language Explanations*. arXiv: 2311.07466 [cs.CL]. URL: https://arxiv.org/abs/2311.07466.

Park, Dong Huk, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach (2018). "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence". In: *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8779–8788. DOI: 10.1109/CVPR.2018.00915.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "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, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "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, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

Pfeiffer, Jonas, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe (July 2022). "Lifting the Curse of Multilinguality by Pre-training Modular Transformers". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 3479–3495. DOI: 10.18653/v1/2022.naacl-main.255. URL: https://aclanthology.org/2022.naacl-main.255.

Piantadosi, Steven T. and Felix Hill (2022). *Meaning without reference in large language models*. arXiv: 2208.02957 [cs.CL].

Pimentel, Tiago, Naomi Saphra, Adina Williams, and Ryan Cotterell (Nov. 2020a). "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, pp. 3138–3153. DOI: 10.18653/v1/2020.emnlp-main.254. URL: https://aclanthology.org/2020.emnlp-main.254.

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell (July 2020b). "Information-Theoretic Probing for Linguistic Structure". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4609–4622. DOI: 10.18653/v1/2020.acl-main.420. URL: https://aclanthology.org/2020.acl-main.420.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding by generative pre-training". In.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *J. Mach. Learn. Res.* 21, 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

Raganato, Alessandro and Jörg Tiedemann (Nov. 2018). "An Analysis of Encoder Representations in Transformer-Based Machine Translation". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 287–297. DOI: 10.18653/v1/W18-5431. URL: https://aclanthology.org/W18-5431.

Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher (July 2019). "Explain Yourself! Leveraging Language Models for Commonsense Reasoning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4932–4942. DOI: 10.18653/v1/P19-1487. URL: https://aclanthology.org/P19-1487.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264.

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg (July 2020). "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, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647. URL: https://aclanthology.org/2020.acl-main.647.

Ravichander, Abhilasha, Yonatan Belinkov, and Eduard Hovy (Apr. 2021). "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, pp. 3363–3377. DOI: 10.18653/v1/2021.eacl-main.295. URL: https://aclanthology.org/2021.eacl-main.295.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

Rissanen, Jorma (1978). "Modeling by shortest data description". In: *Autom.* 14.5, pp. 465–471. DOI: 10.1016/0005-1098(78)90005-5. URL: https://doi.org/10.1016/0005-1098(78)90005-5.

Rogers, Anna, Aleksandr Drozd, and Bofang Li (Aug. 2017). "The (too Many) Problems of Analogical Reasoning with Word Vectors". In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 135–148. DOI: 10.18653/v1/S17-1017. URL: https://aclanthology.org/S17-1017.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://aclanthology.org/2020.tacl-1.54.

Roscher, Ribana, Bastian Bohn, Marco F Duarte, and Jochen Garcke (2020). "Explainable machine learning for scientific insights and discoveries". In: *IEEE Access* 8, pp. 42200–42216.

Ross, Alexis, Matthew Peters, and Ana Marasovic (Dec. 2022). "Does Self-Rationalization Improve Robustness to Spurious Correlations?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7403–7416. DOI: 10.18653/v1/2022.emnlp-main.501. URL: https://aclanthology.org/2022.emnlp-main.501.

Rozanova, Julia, Marco Valentino, Lucas Cordeiro, and André Freitas (May 2023). "Interventional Probing in High Dimensions: An NLI Case Study". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2489–2500. DOI: 10.18653/v1/2023.findings-eacl.188. URL: https://aclanthology.org/2023.findings-eacl.188.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5, pp. 206–215.

Saphra, Naomi and Adam Lopez (June 2019). "Understanding Learning Dynamics Of Language Models with SVCCA". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3257–3267. DOI: 10.18653/v1/N19-1329. URL: https://aclanthology.org/N19-1329.

Serrano, Sofia and Noah A. Smith (July 2019). "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. URL: https://aclanthology.org/P19-1282.

Shen, Hua, Tongshuang Wu, Wenbo Guo, and Ting-Hao Huang (May 2022). "Are Shortest Rationales the Best Explanations for Human Understanding?" In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 10–19. DOI: 10.18653/v1/2022.acl-short.2. URL: https://aclanthology.org/2022.acl-short.2.

Shuster, Kurt, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston (Nov. 2021). "Retrieval Augmentation Reduces Hallucination in Conversation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3784–3803. DOI: 10.18653/v1/2021.findings-emnlp.320. URL: https://aclanthology.org/2021.findings-emnlp.320.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv: 1312.6034 [cs.CV]. URL: https://arxiv.org/abs/1312.6034.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju (2020). "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, pp. 180–186. ISBN: 9781450371100. DOI: 10.1145/3375627.3375830. URL: https://doi.org/10.1145/3375627.3375830.

Slobodkin, Aviv, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel (Dec. 2023). "The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 3607–3625. DOI: 10.18653/v1/2023.emnlp-main.220. URL: https://aclanthology.org/2023.emnlp-main.220.

Stanczak, Karolina, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein (July 2022). "Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1589–1598. DOI: 10.18653/v1/2022.naacl-main.114. URL: https://aclanthology.org/2022.naacl-main.114.

Stańczak, Karolina, Lucas Torroba Hennigen, Adina Williams, Ryan Cotterell, and Isabelle Augenstein (2023). "A latent-variable model for intrinsic probing". In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial*

Intelligence. AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN: 978-1-57735-880-0. DOI: 10.1609/aaai.v37i11.26593. URL: https://doi.org/10.1609/aaai.v37i11.26593.

Sun, Jiuding, Chantal Shaib, and Byron C Wallace (2023). "Evaluating the Zero-shot Robustness of Instruction-tuned Language Models". In: *arXiv preprint arXiv:2306.11270*. URL: https://arxiv.org/pdf/2306.11270.pdf.

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

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Talmor, Alon, Yanai Elazar, Yoav Goldberg, and Jonathan Berant (2020). "oLMpics-On What Language Model Pre-training Captures". In: *Transactions of the Association for Computational Linguistics* 8, pp. 743–758. DOI: 10.1162/tacl_a_00342. URL: https://aclanthology.org/2020.tacl-1.48.

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant (June 2019). "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158. DOI: 10.18653/v1/N19-1421. URL: https://aclanthology.org/N19-1421.

Tan, Chenhao (July 2022). "On the Diversity and Limits of Human Explanations". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2173–2188. DOI: 10.18653/v1/2022.naacl-main.158. URL: https://aclanthology.org/2022.naacl-main.158.

Tedeschi, Simone, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli (July 2023). "What's the Meaning of Superhuman Performance in Today's NLU?" In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 12471–12491. DOI: 10.18653/v1/2023.acl-long.697. URL: https://aclanthology.org/2023.acl-long.697.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick (July 2019). "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. URL: https://aclanthology.org/P19-1452.

Torroba Hennigen, Lucas, Adina Williams, and Ryan Cotterell (Nov. 2020). "Intrinsic Probing through Dimension Selection". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 197–216. DOI: 10.18653/v1/2020.emnlp-main.15. URL: https://aclanthology.org/2020.emnlp-main.15.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (2023). *LLaMA: Open and Efficient Foundation Language Models*. cite arxiv:2302.13971. URL: http://arxiv.org/abs/2302.13971.

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio (July 2010). "Word Representations: A Simple and General Method for Semi-Supervised Learning". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 384–394. URL: https://aclanthology.org/P10-1040.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman (2023). "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting". In: *Thirty-seventh Conference on Neural Information Processing Systems*. New Orleans, Louisiana, USA: Conference on Neural Information Processing Systems (NeurIPS). URL: https://openreview.net/forum?id=bzs4uPLXvi.

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer (2021). "Human evaluation of automatically generated text: Current trends and best practice guidelines". In: *Computer Speech and Language* 67, p. 101151. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2020.101151. URL: https://www.sciencedirect.com/science/article/pii/S088523082030084X.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vig, Jesse and Yonatan Belinkov (Aug. 2019). "Analyzing the Structure of Attention in a Transformer Language Model". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for*

NLP. Florence, Italy: Association for Computational Linguistics, pp. 63–76. DOI: 10.18653/v1/W19-4808. URL: https://aclanthology.org/W19-4808.

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov (July 2019). "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5797–5808. DOI: 10.18653/v1/P19-1580. URL: https://aclanthology.org/P19-1580.

Voita, Elena and Ivan Titov (Nov. 2020). "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, pp. 183–196. DOI: 10.18653/v1/2020.emnlp-main.14. URL: https://aclanthology.org/2020.emnlp-main.14.

Wallace, Eric, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh (Nov. 2019a). "Universal Adversarial Triggers for Attacking and Analyzing NLP". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2153–2162. DOI: 10.18653/v1/D19-1221. URL: https://aclanthology.org/D19-1221.

Wallace, Eric, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh (Nov. 2019b). "AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models". 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): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, pp. 7–12. DOI: 10.18653/v1/D19-3002. URL: https://aclanthology.org/D19-3002.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). "SuperGLUE: a stickier benchmark for general-purpose language understanding systems". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Wang, Cunxiang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang (Dec. 2020). "SemEval-2020 Task 4: Commonsense Validation and Explanation". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 307–321. DOI: 10.18653/v1/2020.semeval-1.39. URL: https://aclanthology.org/2020.semeval-1.39.

Wang, Junlin, Jens Tuyls, Eric Wallace, and Sameer Singh (Nov. 2020). "Gradient-based Analysis of NLP Models is Manipulable". In: *Findings of*

the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, pp. 247–258. DOI: 10.18653/v1/2020.findings-emnlp.24. URL: https://aclanthology.org/2020.findings-emnlp.24.

Wang, Nan, Hongning Wang, Yiling Jia, and Yue Yin (2018). "Explainable Recommendation via Multi-Task Learning in Opinionated Text Data". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, pp. 165–174. ISBN: 9781450356572. DOI: 10.1145/3209978.3210010. URL: https://doi.org/10.1145/3209978.3210010.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2021). "Finetuned Language Models Are Zero-Shot Learners". In: *CoRR* abs/2109.01652. arXiv: 2109.01652. URL: https://arxiv.org/abs/2109.01652.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou (Jan. 2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *arXiv e-prints*, arXiv:2201.11903, arXiv:2201.11903. arXiv: 2201.11903 [cs.CL].

Wiegreffe, Sarah, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi (July 2022). "Reframing Human-AI Collaboration for Generating Free-Text Explanations". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 632–658. DOI: 10.18653/v1/2022.naacl-main.47. URL: https://aclanthology.org/2022.naacl-main.47.

Wiegreffe, Sarah and Ana Marasovic (2021). "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Ed. by J. Vanschoren and S. Yeung. Vol. 1. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/698d51a19d8a121ce581499d7b701668-Paper-round1.pdf.

Wiegreffe, Sarah, Ana Marasović, and Noah A. Smith (Nov. 2021). "Measuring Association Between Labels and Free-Text Rationales". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10266–10284. DOI: 10.18653/v1/2021.emnlp-main.804. URL: https://aclanthology.org/2021.emnlp-main.804.

Wiegreffe, Sarah and Yuval Pinter (Nov. 2019). "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, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://aclanthology.org/D19-1002.

Yu, Mo, Shiyu Chang, Yang Zhang, and Tommi Jaakkola (Nov. 2019). "Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 4094–4103. DOI: 10.18653/v1/D19-1420. URL: https://aclanthology.org/D19-1420.

Zhang, Kelly and Samuel Bowman (Nov. 2018). "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, pp. 359–361. DOI: 10.18653/v1/W18-5448. URL: https://aclanthology.org/W18-5448.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). "BERTScore: Evaluating Text Generation with BERT". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, Yongfeng, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma (2014). "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis". In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14. Gold Coast, Queensland, Australia: Association for Computing Machinery, pp. 83–92. ISBN: 9781450322577. DOI: 10.1145/2600428.2609579. URL: https://doi.org/10.1145/2600428.2609579.

Zhang, Yujia, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell (2019). ""Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations". In: *arXiv preprint arXiv:1904.12991*. URL: https://arxiv.org/pdf/1904.12991.pdf.

Zhao, Xinyan and V. G. Vinod Vydiswaran (2020). "LIREx: Augmenting Language Inference with Relevant Explanation". In: *CoRR* abs/2012.09157. arXiv: 2012.09157. URL: https://arxiv.org/abs/2012.09157.

Zhou, Yangqiaoyu and Chenhao Tan (Nov. 2021). "Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI". In: *Proceedings of the Second Workshop on Insights from Negative Results in NLP*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 117–124. DOI: 10.18653/v1/2021.insights-1.17. URL: https://aclanthology.org/2021.insights-1.17.

Zhu, Yukun, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies

and Reading Books". In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27. URL: https://api.semanticscholar.org/CorpusID:6866988.

Part II

Papers

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

https://doi.org/10.3384/9789180754712

Department of Computer and Information Science
Linköpings universitet

Dissertations

Linköping Studies in Science and Technology
Linköping Studies in Arts and Sciences
Linköping Studies in Statistics
Linköping Studies in Information Science

Linköping Studies in Science and Technology

No 14 **Anders Haraldsson:** A Program Manipulation System Based on Partial Evaluation, 1977, ISBN 91-7372-144-1.

No 17 **Bengt Magnhagen:** Probability Based Verification of Time Margins in Digital Designs, 1977, ISBN 91-7372-157-3.

No 18 **Mats Cedwall**: Semantisk analys av process-beskrivningar i naturligt språk, 1977, ISBN 91- 7372-168-9.

No 22 **Jaak Urmi:** A Machine Independent LISP Compiler and its Implications for Ideal Hardware, 1978, ISBN 91-7372-188-3.

No 33 **Tore Risch:** Compilation of Multiple File Queries in a Meta-Database System, 1978, ISBN 91- 7372-232-4.

No 51 **Erland Jungert:** Synthesizing Database Structures from a User Oriented Data Model, 1980, ISBN 91-7372-387-8.

No 54 **Sture Hägglund:** Contributions to the Development of Methods and Tools for Interactive Design of Applications Software, 1980, ISBN 91-7372-404-1.

No 55 **Pär Emanuelson:** Performance Enhancement in a Well-Structured Pattern Matcher through Partial Evaluation, 1980, ISBN 91-7372-403-3.

No 58 **Bengt Johnsson, Bertil Andersson:** The Human-Computer Interface in Commercial Systems, 1981, ISBN 91-7372-414-9.

No 69 **H. Jan Komorowski:** A Specification of an Abstract Prolog Machine and its Application to Partial Evaluation, 1981, ISBN 91-7372-479-3.

No 71 **René Reboh:** Knowledge Engineering Techniques and Tools for Expert Systems, 1981, ISBN 91-7372-489-0.

No 77 **Östen Oskarsson:** Mechanisms of Modifiability in large Software Systems, 1982, ISBN 91- 7372-527-7.

No 94 **Hans Lunell:** Code Generator Writing Systems, 1983, ISBN 91-7372-652-4.

No 97 **Andrzej Lingas:** Advances in Minimum Weight Triangulation, 1983, ISBN 91-7372-660-5.

No 109 **Peter Fritzson:** Towards a Distributed Programming Environment based on Incremental Compilation, 1984, ISBN 91-7372-801-2.

No 111 **Erik Tengvald:** The Design of Expert Planning Systems. An Experimental Operations Planning System for Turning, 1984, ISBN 91-7372- 805-5.

No 155 **Christos Levcopoulos:** Heuristics for Minimum Decompositions of Polygons, 1987, ISBN 91-7870-133-3.

No 165 **James W. Goodwin:** A Theory and System for Non-Monotonic Reasoning, 1987, ISBN 91-7870-183-X.

No 170 **Zebo Peng:** A Formal Methodology for Automated Synthesis of VLSI Systems, 1987, ISBN 91-7870-225-9.

No 174 **Johan Fagerström:** A Paradigm and System for Design of Distributed Systems, 1988, ISBN 91-7870-301-8.

No 192 **Dimiter Driankov:** Towards a Many Valued Logic of Quantified Belief, 1988, ISBN 91-7870-374-3.

No 213 **Lin Padgham:** Non-Monotonic Inheritance for an Object Oriented Knowledge Base, 1989, ISBN 91-7870-485-5.

No 214 **Tony Larsson:** A Formal Hardware Description and Verification Method, 1989, ISBN 91-7870-517-7.

No 221 **Michael Reinfrank:** Fundamentals and Logical Foundations of Truth Maintenance, 1989, ISBN 91-7870-546-0.

No 239 **Jonas Löwgren:** Knowledge-Based Design Support and Discourse Management in User Interface Management Systems, 1991, ISBN 91-7870-720-X.

No 244 **Henrik Eriksson:** Meta-Tool Support for Knowledge Acquisition, 1991, ISBN 91-7870-746-3.

No 252 **Peter Eklund:** An Epistemic Approach to Interactive Design in Multiple Inheritance Hierarchies, 1991, ISBN 91-7870-784-6.

No 258 **Patrick Doherty:** NML3 - A Non-Monotonic Formalism with Explicit Defaults, 1991, ISBN 91-7870-816-8.

No 260 **Nahid Shahmehri:** Generalized Algorithmic Debugging, 1991, ISBN 91-7870-828-1.

No 264 **Nils Dahlbäck:** Representation of Discourse-Cognitive and Computational Aspects, 1992, ISBN 91-7870-850-8.

No 265 **Ulf Nilsson:** Abstract Interpretations and Abstract Machines: Contributions to a Methodology for the Implementation of Logic Programs, 1992, ISBN 91-7870-858-3.

No 270 **Ralph Rönnquist:** Theory and Practice of Tense-bound Object References, 1992, ISBN 91-7870-873-7.

No 273 **Björn Fjellborg:** Pipeline Extraction for VLSI Data Path Synthesis, 1992, ISBN 91-7870-880-X.

No 276 **Staffan Bonnier:** A Formal Basis for Horn Clause Logic with External Polymorphic Functions, 1992, ISBN 91-7870-896-6.

No 277 **Kristian Sandahl:** Developing Knowledge Management Systems with an Active Expert Methodology, 1992, ISBN 91-7870-897-4.

No 281 **Christer Bäckström:** Computational Complexity of Reasoning about Plans, 1992, ISBN 91-7870-979-2.

No 292 **Mats Wirén:** Studies in Incremental Natural Language Analysis, 1992, ISBN 91-7871-027-8.

No 297 **Mariam Kamkar:** Interprocedural Dynamic Slicing with Applications to Debugging and Testing, 1993, ISBN 91-7871-065-0.

No 302 **Tingting Zhang:** A Study in Diagnosis Using Classification and Defaults, 1993, ISBN 91-7871-078-2.

No 312 **Arne Jönsson:** Dialogue Management for Natural Language Interfaces - An Empirical Approach, 1993, ISBN 91-7871-110-X.

No 338 **Simin Nadjm-Tehrani:** Reactive Systems in Physical Environments: Compositional Modelling and Framework for Verification, 1994, ISBN 91-7871-237-8.

No 371 **Bengt Savén:** Business Models for Decision Support and Learning. A Study of Discrete-Event Manufacturing Simulation at Asea/ABB 1968-1993, 1995, ISBN 91-7871-494-X.

No 375 **Ulf Söderman:** Conceptual Modelling of Mode Switching Physical Systems, 1995, ISBN 91-7871-516-4.

No 383 **Andreas Kågedal:** Exploiting Groundness in Logic Programs, 1995, ISBN 91-7871-538-5.

No 396 **George Fodor:** Ontological Control, Description, Identification and Recovery from Problematic Control Situations, 1995, ISBN 91-7871-603-9.

No 413 **Mikael Pettersson:** Compiling Natural Semantics, 1995, ISBN 91-7871-641-1.

No 414 **Xinli Gu:** RT Level Testability Improvement by Testability Analysis and Transformations, 1996, ISBN 91-7871-654-3.

No 416 **Hua Shu:** Distributed Default Reasoning, 1996, ISBN 91-7871-665-9.

No 429 **Jaime Villegas:** Simulation Supported Industrial Training from an Organisational Learning Perspective - Development and Evaluation of the SSIT Method, 1996, ISBN 91-7871-700-0.

No 431 **Peter Jonsson:** Studies in Action Planning: Algorithms and Complexity, 1996, ISBN 91-7871-704-3.

No 437 **Johan Boye:** Directional Types in Logic Programming, 1996, ISBN 91-7871-725-6.

No 439 **Cecilia Sjöberg:** Activities, Voices and Arenas: Participatory Design in Practice, 1996, ISBN 91-7871-728-0.

No 448 **Patrick Lambrix:** Part-Whole Reasoning in Description Logics, 1996, ISBN 91-7871-820-1.

No 452 **Kjell Orsborn:** On Extensible and Object-Relational Database Technology for Finite Element Analysis Applications, 1996, ISBN 91-7871-827-9.

No 459 **Olof Johansson:** Development Environments for Complex Product Models, 1996, ISBN 91-7871-855-4.

No 461 **Lena Strömbäck:** User-Defined Constructions in Unification-Based Formalisms, 1997, ISBN 91-7871-857-0.

No 462 **Lars Degerstedt:** Tabulation-based Logic Programming: A Multi-Level View of Query Answering, 1996, ISBN 91-7871-858-9.

No 475 **Fredrik Nilsson:** Strategi och ekonomisk styrning - En studie av hur ekonomiska styrsystem utformas och används efter företagsförvärv, 1997, ISBN 91-7871-914-3.

No 480 **Mikael Lindvall:** An Empirical Study of Requirements-Driven Impact Analysis in Object-Oriented Software Evolution, 1997, ISBN 91-7871-927-5.

No 485 **Göran Forslund:** Opinion-Based Systems: The Cooperative Perspective on Knowledge-Based Decision Support, 1997, ISBN 91-7871-938-0.

No 494 **Martin Sköld:** Active Database Management Systems for Monitoring and Control, 1997, ISBN 91-7219-002-7.

No 495 **Hans Olsén:** Automatic Verification of Petri Nets in a CLP framework, 1997, ISBN 91-7219-011-6.

No 498 **Thomas Drakengren:** Algorithms and Complexity for Temporal and Spatial Formalisms, 1997, ISBN 91-7219-019-1.

No 502 **Jakob Axelsson:** Analysis and Synthesis of Heterogeneous Real-Time Systems, 1997, ISBN 91-7219-035-3.

No 503 **Johan Ringström:** Compiler Generation for Data-Parallel Programming Languages from Two-Level Semantics Specifications, 1997, ISBN 91-7219-045-0.

No 512 **Anna Moberg:** Närhet och distans - Studier av kommunikationsmönster i satellitkontor och flexibla kontor, 1997, ISBN 91-7219-119-8.

No 520 **Mikael Ronström:** Design and Modelling of a Parallel Data Server for Telecom Applications, 1998, ISBN 91-7219-169-4.

No 522 **Niclas Ohlsson:** Towards Effective Fault Prevention - An Empirical Study in Software Engineering, 1998, ISBN 91-7219-176-7.

No 526 **Joachim Karlsson:** A Systematic Approach for Prioritizing Software Requirements, 1998, ISBN 91-7219-184-8.

No 530 **Henrik Nilsson:** Declarative Debugging for Lazy Functional Languages, 1998, ISBN 91-7219-197-X.

No 555 **Jonas Hallberg:** Timing Issues in High-Level Synthesis, 1998, ISBN 91-7219-369-7.

No 561 **Ling Lin:** Management of 1-D Sequence Data - From Discrete to Continuous, 1999, ISBN 91-7219-402-2.

No 563 **Eva L Ragnemalm:** Student Modelling based on Collaborative Dialogue with a Learning Companion, 1999, ISBN 91-7219-412-X.

No 567 **Jörgen Lindström:** Does Distance matter? On geographical dispersion in organisations, 1999, ISBN 91-7219-439-1.

No 582 **Vanja Josifovski:** Design, Implementation and Evaluation of a Distributed Mediator System for Data Integration, 1999, ISBN 91-7219-482-0.

No 589 **Rita Kovordányi:** Modeling and Simulating Inhibitory Mechanisms in Mental Image Reinterpretation - Towards Cooperative Human-Computer Creativity, 1999, ISBN 91-7219-506-1.

No 592 **Mikael Ericsson:** Supporting the Use of Design Knowledge - An Assessment of Commenting Agents, 1999, ISBN 91-7219-532-0.

No 593 **Lars Karlsson:** Actions, Interactions and Narratives, 1999, ISBN 91-7219-534-7.

No 594 **C. G. Mikael Johansson:** Social and Organizational Aspects of Requirements Engineering Methods - A practice-oriented approach, 1999, ISBN 91-7219-541-X.

No 595 **Jörgen Hansson:** Value-Driven Multi-Class Overload Management in Real-Time Database Systems, 1999, ISBN 91-7219-542-8.

No 596 **Niklas Hallberg:** Incorporating User Values in the Design of Information Systems and Services in the Public Sector: A Methods Approach, 1999, ISBN 91-7219-543-6.

No 597 **Vivian Vimarlund:** An Economic Perspective on the Analysis of Impacts of Information Technology: From Case Studies in Health-Care towards General Models and Theories, 1999, ISBN 91-7219-544-4.

No 598 **Johan Jenvald:** Methods and Tools in Computer-Supported Taskforce Training, 1999, ISBN 91-7219-547-9.

No 607 **Magnus Merkel:** Understanding and enhancing translation by parallel text processing, 1999, ISBN 91-7219-614-9.

No 611 **Silvia Coradeschi:** Anchoring symbols to sensory data, 1999, ISBN 91-7219-623-8.

No 613 **Man Lin:** Analysis and Synthesis of Reactive Systems: A Generic Layered Architecture Perspective, 1999, ISBN 91-7219-630-0.

No 618 **Jimmy Tjäder:** Systemimplementering i praktiken - En studie av logiker i fyra projekt, 1999, ISBN 91-7219-657-2.

No 627 **Vadim Engelson:** Tools for Design, Interactive Simulation, and Visualization of Object-Oriented Models in Scientific Computing, 2000, ISBN 91-7219-709-9.

No 637 **Esa Falkenroth:** Database Technology for Control and Simulation, 2000, ISBN 91-7219-766-8.

No 639 **Per-Arne Persson:** Bringing Power and Knowledge Together: Information Systems Design for Autonomy and Control in Command Work, 2000, ISBN 91-7219-796-X.

No 660 **Erik Larsson:** An Integrated System-Level Design for Testability Methodology, 2000, ISBN 91-7219-890-7.

No 688 **Marcus Bjäreland:** Model-based Execution Monitoring, 2001, ISBN 91-7373-016-5.

No 689 **Joakim Gustafsson:** Extending Temporal Action Logic, 2001, ISBN 91-7373-017-3.

No 720 **Carl-Johan Petri:** Organizational Information Provision - Managing Mandatory and Discretionary Use of Information Technology, 2001, ISBN 91-7373-126-9.

No 724 **Paul Scerri:** Designing Agents for Systems with Adjustable Autonomy, 2001, ISBN 91-7373-207-9.

No 725 **Tim Heyer:** Semantic Inspection of Software Artifacts: From Theory to Practice, 2001, ISBN 91-7373-208-7.

No 726 **Pär Carlshamre:** A Usability Perspective on Requirements Engineering - From Methodology to Product Development, 2001, ISBN 91-7373-212-5.

No 732 **Juha Takkinen:** From Information Management to Task Management in Electronic Mail, 2002, ISBN 91-7373-258-3.

No 745 **Johan Åberg:** Live Help Systems: An Approach to Intelligent Help for Web Information Systems, 2002, ISBN 91-7373-311-3.

No 746 **Rego Granlund:** Monitoring Distributed Teamwork Training, 2002, ISBN 91-7373-312-1.

No 757 **Henrik André-Jönsson:** Indexing Strategies for Time Series Data, 2002, ISBN 917373-346-6.

No 747 **Anneli Hagdahl:** Development of IT-supported Interorganisational Collaboration - A Case Study in the Swedish Public Sector, 2002, ISBN 91-7373-314-8.

No 749 **Sofie Pilemalm:** Information Technology for Non-Profit Organisations - Extended Participatory Design of an Information System for Trade Union Shop Stewards, 2002, ISBN 91-7373-318-0.

No 765 **Stefan Holmlid:** Adapting users: Towards a theory of use quality, 2002, ISBN 91-7373-397-0.

No 771 **Magnus Morin:** Multimedia Representations of Distributed Tactical Operations, 2002, ISBN 91-7373-421-7.

No 772 **Pawel Pietrzak:** A Type-Based Framework for Locating Errors in Constraint Logic Programs, 2002, ISBN 91-7373-422-5.

No 758 **Erik Berglund:** Library Communication Among Programmers Worldwide, 2002, ISBN 91-7373-349-0.

No 774 **Choong-ho Yi:** Modelling Object-Oriented Dynamic Systems Using a Logic-Based Framework, 2002, ISBN 91-7373-424-1.

No 779 **Mathias Broxvall:** A Study in the Computational Complexity of Temporal Reasoning, 2002, ISBN 91-7373-440-3.

No 793 **Asmus Pandikow:** A Generic Principle for Enabling Interoperability of Structured and Object-Oriented Analysis and Design Tools, 2002, ISBN 91-7373-479-9.

No 785 **Lars Hult:** Publika Informationstjänster. En studie av den Internetbaserade encyklopedins bruksegenskaper, 2003, ISBN 91-7373-461-6.

No 800 **Lars Taxén:** A Framework for the Coordination of Complex Systems´ Development, 2003, ISBN 91-7373-604-X.

No 808 **Klas Gäre:** Tre perspektiv på förväntningar och förändringar i samband med införande av informationssystem, 2003, ISBN 91-7373-618-X.

No 821 **Mikael Kindborg:** Concurrent Comics - programming of social agents by children, 2003, ISBN 91-7373-651-1.

No 823 **Christina Ölvingson:** On Development of Information Systems with GIS Functionality in Public Health Informatics: A Requirements Engineering Approach, 2003, ISBN 91-7373-656-2.

No 828 **Tobias Ritzau:** Memory Efficient Hard Real-Time Garbage Collection, 2003, ISBN 91-7373-666-X.

No 833 **Paul Pop:** Analysis and Synthesis of Communication-Intensive Heterogeneous Real-Time Systems, 2003, ISBN 91-7373-683-X.

No 852 **Johan Moe:** Observing the Dynamic Behaviour of Large Distributed Systems to Improve Development and Testing – An Empirical Study in Software Engineering, 2003, ISBN 91-7373-779-8.

No 867 **Erik Herzog:** An Approach to Systems Engineering Tool Data Representation and Exchange, 2004, ISBN 91-7373-929-4.

No 872 **Aseel Berglund:** Augmenting the Remote Control: Studies in Complex Information Navigation for Digital TV, 2004, ISBN 91-7373-940-5.

No 869 **Jo Skåmedal:** Telecommuting´s Implications on Travel and Travel Patterns, 2004, ISBN 91-7373-935-9.

No 870 **Linda Askenäs:** The Roles of IT - Studies of Organising when Implementing and Using Enterprise Systems, 2004, ISBN 91-7373-936-7.

No 874 **Annika Flycht-Eriksson:** Design and Use of Ontologies in Information-Providing Dialogue Systems, 2004, ISBN 91-7373-947-2.

No 873 **Peter Bunus:** Debugging Techniques for Equation-Based Languages, 2004, ISBN 91-7373-941-3.

No 876 **Jonas Mellin:** Resource-Predictable and Efficient Monitoring of Events, 2004, ISBN 91-7373-956-1.

No 883 **Magnus Bång:** Computing at the Speed of Paper: Ubiquitous Computing Environments for Healthcare Professionals, 2004, ISBN 91-7373-971-5.

No 882 **Robert Eklund:** Disfluency in Swedish human-human and human-machine travel booking dialogues, 2004, ISBN 91-7373-966-9.

No 887 **Anders Lindström:** English and other Foreign Linguistic Elements in Spoken Swedish. Studies of Productive Processes and their Modelling using Finite-State Tools, 2004, ISBN 91-7373-981-2.

No 889 **Zhiping Wang:** Capacity-Constrained Production-inventory systems - Modelling and Analysis in both a traditional and an e-business context, 2004, ISBN 91-85295-08-6.

No 893 **Pernilla Qvarfordt:** Eyes on Multimodal Interaction, 2004, ISBN 91-85295-30-2.

No 910 **Magnus Kald:** In the Borderland between Strategy and Management Control - Theoretical Framework and Empirical Evidence, 2004, ISBN 91-85295-82-5.

No 918 **Jonas Lundberg:** Shaping Electronic News: Genre Perspectives on Interaction Design, 2004, ISBN 91-85297-14-3.

No 900 **Mattias Arvola:** Shades of use: The dynamics of interaction design for sociable use, 2004, ISBN 91-85295-42-6.

No 920 **Luis Alejandro Cortés:** Verification and Scheduling Techniques for Real-Time Embedded Systems, 2004, ISBN 91-85297-21-6.

No 929 **Diana Szentivanyi:** Performance Studies of Fault-Tolerant Middleware, 2005, ISBN 91-85297-58-5.

No 933 **Mikael Cäker:** Management Accounting as Constructing and Opposing Customer Focus: Three Case Studies on Management Accounting and Customer Relations, 2005, ISBN 91-85297-64-X.

No 937 **Jonas Kvarnström:** TALplanner and Other Extensions to Temporal Action Logic, 2005, ISBN 91-85297-75-5.

No 938 **Bourhane Kadmiry:** Fuzzy Gain-Scheduled Visual Servoing for Unmanned Helicopter, 2005, ISBN 91-85297-76-3.

No 945 **Gert Jervan:** Hybrid Built-In Self-Test and Test Generation Techniques for Digital Systems, 2005, ISBN 91-85297-97-6.

No 946 **Anders Arpteg:** Intelligent Semi-Structured Information Extraction, 2005, ISBN 91-85297-98-4.

No 947 **Ola Angelsmark:** Constructing Algorithms for Constraint Satisfaction and Related Problems - Methods and Applications, 2005, ISBN 91-85297-99-2.

No 963 **Calin Curescu:** Utility-based Optimisation of Resource Allocation for Wireless Networks, 2005, ISBN 91-85457-07-8.

No 972 **Björn Johansson:** Joint Control in Dynamic Situations, 2005, ISBN 91-85457-31-0.

No 974 **Dan Lawesson:** An Approach to Diagnosability Analysis for Interacting Finite State Systems, 2005, ISBN 91-85457-39-6.

No 979 **Claudiu Duma:** Security and Trust Mechanisms for Groups in Distributed Services, 2005, ISBN 91-85457-54-X.

No 983 **Sorin Manolache:** Analysis and Optimisation of Real-Time Systems with Stochastic Behaviour, 2005, ISBN 91-85457-60-4.

No 986 **Yuxiao Zhao:** Standards-Based Application Integration for Business-to-Business Communications, 2005, ISBN 91-85457-66-3.

No 1004 **Patrik Haslum:** Admissible Heuristics for Automated Planning, 2006, ISBN 91-85497-28-2.

No 1005 **Aleksandra Tešanovic:** Developing Reusable and Reconfigurable Real-Time Software using Aspects and Components, 2006, ISBN 91-85497-29-0.

No 1008 **David Dinka:** Role, Identity and Work: Extending the design and development agenda, 2006, ISBN 91-85497-42-8.

No 1009 **Iakov Nakhimovski:** Contributions to the Modeling and Simulation of Mechanical Systems with Detailed Contact Analysis, 2006, ISBN 91-85497-43-X.

No 1013 **Wilhelm Dahllöf:** Exact Algorithms for Exact Satisfiability Problems, 2006, ISBN 91-85523-97-6.

No 1016 **Levon Saldamli:** PDEModelica - A High-Level Language for Modeling with Partial Differential Equations, 2006, ISBN 91-85523-84-4.

No 1017 **Daniel Karlsson:** Verification of Component-based Embedded System Designs, 2006, ISBN 91-85523-79-8

No 1018 **Ioan Chisalita:** Communication and Networking Techniques for Traffic Safety Systems, 2006, ISBN 91-85523-77-1.

No 1019 **Tarja Susi:** The Puzzle of Social Activity - The Significance of Tools in Cognition and Cooperation, 2006, ISBN 91-85523-71-2.

No 1021 **Andrzej Bednarski:** Integrated Optimal Code Generation for Digital Signal Processors, 2006, ISBN 91-85523-69-0.

No 1022 **Peter Aronsson:** Automatic Parallelization of Equation-Based Simulation Programs, 2006, ISBN 91-85523-68-2.

No 1030 **Robert Nilsson:** A Mutation-based Framework for Automated Testing of Timeliness, 2006, ISBN 91-85523-35-6.

No 1034 **Jon Edvardsson:** Techniques for Automatic Generation of Tests from Programs and Specifications, 2006, ISBN 91-85523-31-3.

No 1035 **Vaida Jakoniene:** Integration of Biological Data, 2006, ISBN 91-85523-28-3.

No 1045 **Genevieve Gorrell:** Generalized Hebbian Algorithms for Dimensionality Reduction in Natural Language Processing, 2006, ISBN 91-85643-88-2.

No 1051 **Yu-Hsing Huang:** Having a New Pair of Glasses - Applying Systemic Accident Models on Road Safety, 2006, ISBN 91-85643-64-5.

No 1054 **Åsa Hedenskog:** Perceive those things which cannot be seen - A Cognitive Systems Engineering perspective on requirements management, 2006, ISBN 91-85643-57-2.

No 1061 **Cécile Åberg:** An Evaluation Platform for Semantic Web Technology, 2007, ISBN 91-85643-31-9.

No 1073 **Mats Grindal:** Handling Combinatorial Explosion in Software Testing, 2007, ISBN 978-91-85715-74-9.

No 1075 **Almut Herzog:** Usable Security Policies for Runtime Environments, 2007, ISBN 978-91-85715-65-7.

No 1079 **Magnus Wahlström:** Algorithms, measures, and upper bounds for Satisfiability and related problems, 2007, ISBN 978-91-85715-55-8.

No 1083 **Jesper Andersson:** Dynamic Software Architectures, 2007, ISBN 978-91-85715-46-6.

No 1086 **Ulf Johansson:** Obtaining Accurate and Comprehensible Data Mining Models - An Evolutionary Approach, 2007, ISBN 978-91-85715-34-3.

No 1089 **Traian Pop:** Analysis and Optimisation of Distributed Embedded Systems with Heterogeneous Scheduling Policies, 2007, ISBN 978-91-85715-27-5.

No 1091 **Gustav Nordh:** Complexity Dichotomies for CSP-related Problems, 2007, ISBN 978-91-85715-20-6.

No 1106 **Per Ola Kristensson:** Discrete and Continuous Shape Writing for Text Entry and Control, 2007, ISBN 978-91-85831-77-7.

No 1110 **He Tan:** Aligning Biomedical Ontologies, 2007, ISBN 978-91-85831-56-2.

No 1112 **Jessica Lindblom:** Minding the body - Interacting socially through embodied action, 2007, ISBN 978-91-85831-48-7.

No 1113 **Pontus Wärnestål:** Dialogue Behavior Management in Conversational Recommender Systems, 2007, ISBN 978-91-85831-47-0.

No 1120 **Thomas Gustafsson:** Management of Real-Time Data Consistency and Transient Overloads in Embedded Systems, 2007, ISBN 978-91-85831-33-3.

No 1127 **Alexandru Andrei:** Energy Efficient and Predictable Design of Real-time Embedded Systems, 2007, ISBN 978-91-85831-06-7.

No 1139 **Per Wikberg:** Eliciting Knowledge from Experts in Modeling of Complex Systems: Managing Variation and Interactions, 2007, ISBN 978-91-85895-66-3.

No 1143 **Mehdi Amirijoo:** QoS Control of Real-Time Data Services under Uncertain Workload, 2007, ISBN 978-91-85895-49-6.

No 1150 **Sanny Syberfeldt:** Optimistic Replication with Forward Conflict Resolution in Distributed Real-Time Databases, 2007, ISBN 978-91-85895-27-4.

No 1155 **Beatrice Alenljung:** Envisioning a Future Decision Support System for Requirements Engineering - A Holistic and Human-centred Perspective, 2008, ISBN 978-91-85895-11-3.

No 1156 **Artur Wilk:** Types for XML with Application to Xcerpt, 2008, ISBN 978-91-85895-08-3.

No 1183 **Adrian Pop:** Integrated Model-Driven Development Environments for Equation-Based Object-Oriented Languages, 2008, ISBN 978-91-7393-895-2.

No 1185 **Jörgen Skågeby:** Gifting Technologies - Ethnographic Studies of End-users and Social Media Sharing, 2008, ISBN 978-91-7393-892-1.

No 1187 **Imad-Eldin Ali Abugessaisa:** Analytical tools and information-sharing methods supporting road safety organizations, 2008, ISBN 978-91-7393-887-7.

No 1204 **H. Joe Steinhauer:** A Representation Scheme for Description and Reconstruction of Object Configurations Based on Qualitative Relations, 2008, ISBN 978-91-7393-823-5.

No 1222 **Anders Larsson:** Test Optimization for Core-based System-on-Chip, 2008, ISBN 978-91-7393-768-9.

No 1238 **Andreas Borg:** Processes and Models for Capacity Requirements in Telecommunication Systems, 2009, ISBN 978-91-7393-700-9.

No 1240 **Fredrik Heintz:** DyKnow: A Stream-Based Knowledge Processing Middleware Framework, 2009, ISBN 978-91-7393-696-5.

No 1241 **Birgitta Lindström:** Testability of Dynamic Real-Time Systems, 2009, ISBN 978-91-7393-695-8.

No 1244 **Eva Blomqvist:** Semi-automatic Ontology Construction based on Patterns, 2009, ISBN 978-91-7393-683-5.

No 1249 **Rogier Woltjer:** Functional Modeling of Constraint Management in Aviation Safety and Command and Control, 2009, ISBN 978-91-7393-659-0.

No 1260 **Gianpaolo Conte:** Vision-Based Localization and Guidance for Unmanned Aerial Vehicles, 2009, ISBN 978-91-7393-603-3.

No 1262 **AnnMarie Ericsson:** Enabling Tool Support for Formal Analysis of ECA Rules, 2009, ISBN 978-91-7393-598-2.

No 1266 **Jiri Trnka:** Exploring Tactical Command and Control: A Role-Playing Simulation Approach, 2009, ISBN 978-91-7393-571-5.

No 1268 **Bahlol Rahimi:** Supporting Collaborative Work through ICT - How End-users Think of and Adopt Integrated Health Information Systems, 2009, ISBN 978-91-7393-550-0.

No 1274 **Fredrik Kuivinen:** Algorithms and Hardness Results for Some Valued CSPs, 2009, ISBN 978-91-7393-525-8.

No 1281 **Gunnar Mathiason:** Virtual Full Replication for Scalable Distributed Real-Time Databases, 2009, ISBN 978-91-7393-503-6.

No 1290 **Viacheslav Izosimov:** Scheduling and Optimization of Fault-Tolerant Distributed Embedded Systems, 2009, ISBN 978-91-7393-482-4.

No 1294 **Johan Thapper:** Aspects of a Constraint Optimisation Problem, 2010, ISBN 978-91-7393-464-0.

No 1306 **Susanna Nilsson:** Augmentation in the Wild: User Centered Development and Evaluation of Augmented Reality Applications, 2010, ISBN 978-91-7393-416-9.

No 1313 **Christer Thörn:** On the Quality of Feature Models, 2010, ISBN 978-91-7393-394-0.

No 1321 **Zhiyuan He:** Temperature Aware and Defect-Probability Driven Test Scheduling for System-on-Chip, 2010, ISBN 978-91-7393-378-0.

No 1333 **David Broman:** Meta-Languages and Semantics for Equation-Based Modeling and Simulation, 2010, ISBN 978-91-7393-335-3.

No 1337 **Alexander Siemers:** Contributions to Modelling and Visualisation of Multibody Systems Simulations with Detailed Contact Analysis, 2010, ISBN 978-91-7393-317-9.

No 1354 **Mikael Asplund:** Disconnected Discoveries: Availability Studies in Partitioned Networks, 2010, ISBN 978-91-7393-278-3.

No 1359 **Jana Rambusch**: Mind Games Extended: Understanding Gameplay as Situated Activity, 2010, ISBN 978-91-7393-252-3.

No 1373 **Sonia Sangari**: Head Movement Correlates to Focus Assignment in Swedish, 2011, ISBN 978-91-7393-154-0.

No 1374 **Jan-Erik Källhammer**: Using False Alarms when Developing Automotive Active Safety Systems, 2011, ISBN 978-91-7393-153-3.

No 1375 **Mattias Eriksson**: Integrated Code Generation, 2011, ISBN 978-91-7393-147-2.

No 1381 **Ola Leifler**: Affordances and Constraints of Intelligent Decision Support for Military Command and Control – Three Case Studies of Support Systems, 2011, ISBN 978-91-7393-133-5.

No 1386 **Soheil Samii**: Quality-Driven Synthesis and Optimization of Embedded Control Systems, 2011, ISBN 978-91-7393-102-1.

No 1419 **Erik Kuiper**: Geographic Routing in Intermittently-connected Mobile Ad Hoc Networks: Algorithms and Performance Models, 2012, ISBN 978-91-7519-981-8.

No 1451 **Sara Stymne**: Text Harmonization Strategies for Phrase-Based Statistical Machine Translation, 2012, ISBN 978-91-7519-887-3.

No 1455 **Alberto Montebelli**: Modeling the Role of Energy Management in Embodied Cognition, 2012, ISBN 978-91-7519-882-8.

No 1465 **Mohammad Saifullah**: Biologically-Based Interactive Neural Network Models for Visual Attention and Object Recognition, 2012, ISBN 978-91-7519-838-5.

No 1490 **Tomas Bengtsson**: Testing and Logic Optimization Techniques for Systems on Chip, 2012, ISBN 978-91-7519-742-5.

No 1481 **David Byers**: Improving Software Security by Preventing Known Vulnerabilities, 2012, ISBN 978-91-7519-784-5.

No 1496 **Tommy Färnqvist**: Exploiting Structure in CSP-related Problems, 2013, ISBN 978-91-7519-711-1.

No 1503 **John Wilander**: Contributions to Specification, Implementation, and Execution of Secure Software, 2013, ISBN 978-91-7519-681-7.

No 1506 **Magnus Ingmarsson**: Creating and Enabling the Useful Service Discovery Experience, 2013, ISBN 978-91-7519-662-6.

No 1547 **Wladimir Schamai**: Model-Based Verification of Dynamic System Behavior against Requirements: Method, Language, and Tool, 2013, ISBN 978-91-7519-505-6.

No 1551 **Henrik Svensson**: Simulations, 2013, ISBN 978-91-7519-491-2.

No 1559 **Sergiu Rafiliu**: Stability of Adaptive Distributed Real-Time Systems with Dynamic Resource Management, 2013, ISBN 978-91-7519-471-4.

No 1581 **Usman Dastgeer**: Performance-aware Component Composition for GPU-based Systems, 2014, ISBN 978-91-7519-383-0.

No 1602 **Cai Li**: Reinforcement Learning of Locomotion based on Central Pattern Generators, 2014, ISBN 978-91-7519-313-7.

No 1652 **Roland Samlaus**: An Integrated Development Environment with Enhanced Domain-Specific Interactive Model Validation, 2015, ISBN 978-91-7519-090-7.

No 1663 **Hannes Uppman**: On Some Combinatorial Optimization Problems: Algorithms and Complexity, 2015, ISBN 978-91-7519-072-3.

No 1664 **Martin Sjölund**: Tools and Methods for Analysis, Debugging, and Performance Improvement of Equation-Based Models, 2015, ISBN 978-91-7519-071-6.

No 1666 **Kristian Stavåker**: Contributions to Simulation of Modelica Models on Data-Parallel Multi-Core Architectures, 2015, ISBN 978-91-7519-068-6.

No 1680 **Adrian Lifa**: Hardware/Software Codesign of Embedded Systems with Reconfigurable and Heterogeneous Platforms, 2015, ISBN 978-91-7519-040-2.

No 1685 **Bogdan Tanasa**: Timing Analysis of Distributed Embedded Systems with Stochastic Workload and Reliability Constraints, 2015, ISBN 978-91-7519-022-8.

No 1691 **Håkan Warnquist**: Troubleshooting Trucks – Automated Planning and Diagnosis, 2015, ISBN 978-91-7685-993-3.

No 1702 **Nima Aghaee**: Thermal Issues in Testing of Advanced Systems on Chip, 2015, ISBN 978-91-7685-949-0.

No 1715 **Maria Vasilevskaya**: Security in Embedded Systems: A Model-Based Approach with Risk Metrics, 2015, ISBN 978-91-7685-917-9.

No 1729 **Ke Jiang**: Security-Driven Design of Real-Time Embedded System, 2016, ISBN 978-91-7685-884-4.

No 1733 **Victor Lagerkvist**: Strong Partial Clones and the Complexity of Constraint Satisfaction Problems: Limitations and Applications, 2016, ISBN 978-91-7685-856-1.

No 1734 **Chandan Roy**: An Informed System Development Approach to Tropical Cyclone Track and Intensity Forecasting, 2016, ISBN 978-91-7685-854-7.

No 1746 **Amir Aminifar**: Analysis, Design, and Optimization of Embedded Control Systems, 2016, ISBN 978-91-7685-826-4.

No 1747 **Ekhiotz Vergara**: Energy Modelling and Fairness for Efficient Mobile Communication, 2016, ISBN 978-91-7685-822-6.

No 1748 **Dag Sonntag**: Chain Graphs – Interpretations, Expressiveness and Learning Algorithms, 2016, ISBN 978-91-7685-818-9.

No 1768 **Anna Vapen**: Web Authentication using Third-Parties in Untrusted Environments, 2016, ISBN 978-91-7685-753-3.

No 1778 **Magnus Jandinger**: On a Need to Know Basis: A Conceptual and Methodological Framework for Modelling and Analysis of Information Demand in an Enterprise Context, 2016, ISBN 978-91-7685-713-7.

No 1798 **Rahul Hiran**: Collaborative Network Security: Targeting Wide-area Routing and Edge-network Attacks, 2016, ISBN 978-91-7685-662-8.

No 1813 **Nicolas Melot**: Algorithms and Framework for Energy Efficient Parallel Stream Computing on Many-Core Architectures, 2016, ISBN 978-91-7685-623-9.

No 1823 **Amy Rankin**: Making Sense of Adaptations: Resilience in High-Risk Work, 2017, ISBN 978-91-7685-596-6.

No 1831 **Lisa Malmberg**: Building Design Capability in the Public Sector: Expanding the Horizons of Development, 2017, ISBN 978-91-7685-585-0.

No 1851 **Marcus Bendtsen**: Gated Bayesian Networks, 2017, ISBN 978-91-7685-525-6.

No 1852 **Zlatan Dragisic**: Completion of Ontologies and Ontology Networks, 2017, ISBN 978-91-7685-522-5.

No 1854 **Meysam Aghighi**: Computational Complexity of some Optimization Problems in Planning, 2017, ISBN 978-91-7685-519-5.

No 1863 **Simon Ståhlberg**: Methods for Detecting Unsolvable Planning Instances using Variable Projection, 2017, ISBN 978-91-7685-498-3.

No 1879 **Karl Hammar**: Content Ontology Design Patterns: Qualities, Methods, and Tools, 2017, ISBN 978-91-7685-454-9.

No 1887 **Ivan Ukhov**: System-Level Analysis and Design under Uncertainty, 2017, ISBN 978-91-7685-426-6.

No 1891 **Valentina Ivanova**: Fostering User Involvement in Ontology Alignment and Alignment Evaluation, 2017, ISBN 978-91-7685-403-7.

No 1902 **Vengatanathan Krishnamoorthi**: Efficient HTTP-based Adaptive Streaming of Linear and Interactive Videos, 2018, ISBN 978-91-7685-371-9.

No 1903 **Lu Li**: Programming Abstractions and Optimization Techniques for GPU-based Heterogeneous Systems, 2018, ISBN 978-91-7685-370-2.

No 1913 **Jonas Rybing**: Studying Simulations with Distributed Cognition, 2018, ISBN 978-91-7685-348-1.

No 1936 **Leif Jonsson**: Machine Learning-Based Bug Handling in Large-Scale Software Development, 2018, ISBN 978-91-7685-306-1.

No 1964 **Arian Maghazeh**: System-Level Design of GPU-Based Embedded Systems, 2018, ISBN 978-91-7685-175-3.

No 1967 **Mahder Gebremedhin**: Automatic and Explicit Parallelization Approaches for Equation Based Mathematical Modeling and Simulation, 2019, ISBN 978-91-7685-163-0.

No 1984 **Anders Andersson**: Distributed Moving Base Driving Simulators – Technology, Performance, and Requirements, 2019, ISBN 978-91-7685-090-9.

No 1993 **Ulf Kargén**: Scalable Dynamic Analysis of Binary Code, 2019, ISBN 978-91-7685-049-7.

No 2001 **Tim Overkamp**: How Service Ideas Are Implemented: Ways of Framing and Addressing Service Transformation, 2019, ISBN 978-91-7685-025-1.

No 2006 **Daniel de Leng**: Robust Stream Reasoning Under Uncertainty, 2019, ISBN 978-91-7685-013-8.

No 2048 **Biman Roy**: Applications of Partial Polymorphisms in (Fine-Grained) Complexity of Constraint Satisfaction Problems, 2020, ISBN 978-91-7929-898-2.

No 2051 **Olov Andersson**: Learning to Make Safe Real-Time Decisions Under Uncertainty for Autonomous Robots, 2020, ISBN 978-91-7929-889-0.

No 2065 **Vanessa Rodrigues**: Designing for Resilience: Navigating Change in Service Systems, 2020, ISBN 978-91-7929-867-8.

No 2082 **Robin Kurtz**: Contributions to Semantic Dependency Parsing: Search, Learning, and Application, 2020, ISBN 978-91-7929-822-7.

No 2108 **Shanai Ardi**: Vulnerability and Risk Analysis Methods and Application in Large Scale Development of Secure Systems, 2021, ISBN 978-91-7929-744-2.

No 2125 **Zeinab Ganjei**: Parameterized Verification of Synchronized Concurrent Programs, 2021, ISBN 978-91-7929-697-1.

No 2153 **Robin Keskisärkkä**: Complex Event Processing under Uncertainty in RDF Stream Processing, 2021, ISBN 978-91-7929-621-6.

No 2168 **Rouhollah Mahfouzi**: Security-Aware Design of Cyber-Physical Systems for Control Applications, 2021, ISBN 978-91-7929-021-4.

No 2205 **August Ernstsson**: Pattern-based Programming Abstractions for Heterogeneous Parallel Computing, 2022, ISBN 978-91-7929-195-2.

No 2218 **Huanyu Li**: Ontology-Driven Data Access and Data Integration with an Application in the Materials Design Domain, 2022, ISBN 978-91-7929-267-6.

No 2219 **Evelina Rennes**: Automatic Adaption of Swedish Text for Increased Inclusion, 2022, ISBN 978-91-7929-269-0.

No 2220 **Yuanbin Zhou**: Synthesis of Safety-Critical Real-Time Systems, 2022, ISBN 978-91-7929-271-3.

No 2247 **Azeem Ahmad**: Contributions to Improving Feedback and Trust in Automated Testing and Continuous Integration and Delivery, 2022, ISBN 978-91-7929-422-9.

No 2248 **Ana Kuštrak Korper**: Innovating Innovation: Understanding the Role of Service Design in Service Innovation, 2022, ISBN 978-91-7929-424-3.

No 2256 **Adrian Horga:** Performance and Security Analysis for GPU-Based Applications, 2022, ISBN 978-91-7929-487-8.

No 2262 **Mattias Tiger:** Safety-Aware Autonomous Systems: Preparing Robots for Life in the Real World 2022, ISBN 978-91-7929-501-1.

No 2266 **Chih-Yuan Lin:** Network-based Anomaly Detection for SCADA Systems: Traffic Generation and Modeling, 2022, ISBN 978-91-7929-517-2.

No 2280 **Filip Strömbäck:** Teaching and Learning Concurrent Programming in the Shared Memory Model, 2023, ISBN 978-91-8075-000-4.

No 2298 **Fiona Lambe:** Devising Capabilities: Service Design for Development Interventions, 2023, ISBN 978-91-8075-080-6.

No 2309 **Alachew Mengist:** Model-Based Tools Integration and Ontology-Driven Traceability in Model-Based Development Environments, 2023, ISBN 978-91-8075-143-8.

No 2322 **Mariusz Wzorek:** Selected Functionalities for Autonomous Intelligent Systems in Public Safety Scenarios, 2023, ISBN 978-91-8075-195-7.

No 2351 **Johan Källström:** Reinforcement Learning for Improved Utility of Simulation-Based Training, 2023, ISBN 978-91-8075-366-1.

No 2364 **Jenny Kunz:** Understanding Large Language Models: Towards Rigorous and Targeted Interpretability Using Probing Classifiers and Self-Rationalisation, 2024, ISBN 978-91-8075-470-5.

No 2366 **Sijin Cheng:** Query Processing over Heterogeneous Federations of Graph Data, 2024, ISBN 978-91-8075-488-0.

Linköping Studies in Arts and Sciences

No 504 **Ing-Marie Jonsson:** Social and Emotional Characteristics of Speech-based In-Vehicle Information Systems: Impact on Attitude and Driving Behaviour, 2009, ISBN 978-91-7393-478-7.

No 586 **Fabian Segelström:** Stakeholder Engagement for Service Design: How service designers identify and communicate insights, 2013, ISBN 978-91-7519-554-4.

No 618 **Johan Blomkvist:** Representing Future Situations of Service: Prototyping in Service Design, 2014, ISBN 978-91-7519-343-4.

No 620 **Marcus Mast:** Human-Robot Interaction for Semi-Autonomous Assistive Robots, 2014, ISBN 978-91-7519-319-9.

No 677 **Peter Berggren:** Assessing Shared Strategic Understanding, 2016, ISBN 978-91-7685-786-1.

No 695 **Mattias Forsblad:** Distributed cognition in home environments: The prospective memory and cognitive practices of older adults, 2016, ISBN 978-91-7685-686-4.

No 787 **Sara Nygårdhs:** Adaptive behaviour in traffic: An individual road user perspective, 2020, ISBN 978-91-7929-857-9.

No 811 **Sam Thellman:** Social Robots as Intentional Agents, 2021, ISBN, 978-91-7929-008-5.

Linköping Studies in Statistics

No 9 **Davood Shahsavani:** Computer Experiments Designed to Explore and Approximate Complex Deterministic Models, 2008, ISBN 978-91-7393-976-8.

No 10 **Karl Wahlin:** Roadmap for Trend Detection and Assessment of Data Quality, 2008, ISBN 978-91-7393-792-4.

No 11 **Oleg Sysoev:** Monotonic regression for large multivariate datasets, 2010, ISBN 978-91-7393-412-1.

No 13 **Agné Burauskaite-Harju:** Characterizing Temporal Change and Inter-Site Correlations in Daily and Subdaily Precipitation Extremes, 2011, ISBN 978-91-7393-110-6.

No 14 **Måns Magnusson:** Scalable and Efficient Probabilistic Topic Model Inference for Textual Data, 2018, ISBN 978-91-7685-288-0.

No 15 **Per Sidén:** Scalable Bayesian spatial analysis with Gaussian Markov random fields, 2020, 978-91-7929-818-0.

No 16 **Caroline Svahn:** Prediction Methods for High Dimensional Data with Censored Covariates, 2022, 978-91-7929-398-7.

No 17 **Héctor Rodriguez Déniz:** Bayesian Models for Spatiotemporal Data from Transportation Networks, 2023, 978-91-8075-035-6.

Linköping Studies in Information Science

No 1 **Karin Axelsson:** Metodisk systemstrukturering- att skapa samstämmighet mellan informationssystem-arkitektur och verksamhet, 1998. ISBN 9172-19-296-8.

No 2 **Stefan Cronholm:** Metodverktyg och användbarhet - en studie av datorstödd metodbaserad systemutveckling, 1998, ISBN 9172-19-299-2.

No 3 **Anders Avdic:** Användare och utvecklare - om anveckling med kalkylprogram, 1999. ISBN 91-7219-606-8.

No 4 **Owen Eriksson:** Kommunikationskvalitet hos informationssystem och affärsprocesser, 2000, ISBN 91-7219-811-7.

No 5 **Mikael Lind:** Från system till process - kriterier för processbestämning vid verksamhetsanalys, 2001, ISBN 91-7373-067-X.

No 6 **Ulf Melin:** Koordination och informationssystem i företag och nätverk, 2002, ISBN 91-7373-278-8.

No 7 **Pär J. Ågerfalk:** Information Systems Actability - Understanding Information Technology as a Tool for Business Action and Communication, 2003, ISBN 91-7373-628-7.

No 8 **Ulf Seigerroth:** Att förstå och förändra system-utvecklingsverksamheter - en taxonomi för metautveckling, 2003, ISBN 91-7373-736-4.

No 9 **Karin Hedström:** Spår av datoriseringens värden – Effekter av IT i äldreomsorg, 2004, ISBN 91-7373-963-4.

No 10 **Ewa Braf:** Knowledge Demanded for Action - Studies on Knowledge Mediation in Organisations, 2004, ISBN 91-85295-47-7.

No 11 **Fredrik Karlsson:** Method Configuration method and computerized tool support, 2005, ISBN 91-85297-48-8.

No 12 **Malin Nordström:** Styrbar systemförvaltning - Att organisera systemförvaltningsverksamhet med hjälp av effektiva förvaltningsobjekt, 2005, ISBN 91-85297-60-7.

No 13 **Stefan Holgersson:** Yrke: POLIS - Yrkeskunskap, motivation, IT-system och andra förutsättningar för polisarbete, 2005, ISBN 91-85299-43-X.

No 14 **Benneth Christiansson, Marie-Therese Christiansson:** Mötet mellan process och komponent - mot ett ramverk för en verksamhetsnära kravspecifikation vid anskaffning av komponent-baserade informationssystem, 2006, ISBN 91-85643-22-X.

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2364, 2024
Department of Computer and Information Science

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se