Thesis for The Degree of Doctor of Philosophy

Fact and Ideology in the Machine: Modelling Knowledge and Belief in Neural Models from Text

Denitsa Saynova

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2025

Fact and Ideology in the Machine: Modelling Knowledge and Belief in Neural Models from Text

Denitsa Saynova

ISBN 978-91-8103-272-7
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5730.
ISSN 0346-718X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2025.

To my family and friends.

Fact and Ideology in the Machine: Modelling Knowledge and Belief in Neural Models from Text

Denitsa Saynova

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

This thesis explores questions of knowledge, language, and neural network models. Motivated by an increasing need for insight into complex political and social science phenomena, we study how methods within natural language processing (NLP) can help us gain such insight. With a particular focus on a model's knowledge, how it is structured, and how we can access and assess it, we study two important aspects of NLP models.

First, we investigate their capabilities and limitations, focusing on how they can capture political and social signals. We use embedding models to capture and reveal distinctions in policy and ideology in Swedish political parties, discussing the strengths and drawbacks of the approach. We also investigate the presence of more complex social knowledge in large pre-trained language models. We prompt models to produce synthetic samples of responses to social science experiments and access if effects calculated from the synthetic data can be used to predict a study's replicability. A central limitation we find in these studies is the lack of robustness, which we explore in depth by studying what influences model consistency in a more simplified setting – namely – recalling facts.

Second, we aim to bridge the gap between the model and the domain expert by developing and improving interpretability insights of model behaviour. We develop a method for aggregating class-level explanations for a text classifier and demonstrate its utility in the context of Swedish political texts. We also develop the understanding of how models store and access factual information. We propose a taxonomy of possible language model behaviours for fact completion and, based on our novel testing data set, examine internal knowledge structures using established mechanistic interpretability methods.

Keywords

Natural Language Processing, Political Science, Representation, Explainability, Evaluation

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] A. Fredén, M. Johansson, **D. Saynova**
Word embeddings on ideology and issues from Swedish parliamentarians' motions: a comparative approach
Journal of Elections, Public Opinion and Parties (Dec 2024), 1–22.

[**Paper II**] **D. Saynova**, B. Bruinsma, M. Johansson, R. Johansson
Class Explanations: the Role of Domain-Specific Content and Stop Words
In Proceedings of the 24th Nordic Conference on Computational Linguistics (2023).

[**Paper III**] L. Hagström, **D. Saynova**, T. Norlund, M. Johansson, R. Johansson
The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models
In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.

[**Paper IV**] **D. Saynova**[*] , L. Hagström[*], M. Johansson, R. Johansson, M. Kuhlmann
Fact Recall, Heuristics or Pure Guesswork? Precise Interpretations of Language Models for Fact Completion
forthcoming, In Findings of the Association for Computational Linguistics: ACL 2025.

[**Paper V**] **D. Saynova**[*], K. Hansson[*], B. Bruinsma, A. Fredén, M. Johansson
Identifying Non-Replicable Social Science Studies with Language Models
abstract accepted at 11th International Conference on Computational Social Science (2025).

[*]Equal contribution.

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] B. Bruinsma, A. Fredén, K. Hansson, M. Johansson, P. Kisić-Merino, **Denitsa Saynova**
Setting the AI Agenda – Evidence from Sweden in the ChatGPT Era
AEQUITAS 2024: Workshop on Fairness and Bias in AI, co-located with ECAI 2024.

[**b**] B. Bonafilia, B. Bruinsma, **Denitsa Saynova**, M. Johansson
Sudden Semantic Shifts in Swedish NATO discourse
In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), p 184–193.

Acknowledgment

I would like to thank my family and friends. To Blago and Vihra, for their constant support without judgement even at my most difficult moments. To Albena, Vanya, Lyudmil, Sophia and Olivier – thank you for your guidance and love and giving me so much joy. To Diya and Iva – thank you for keeping me sane, and to David and Igor – our chats have been a true source of comfort.

To my supervisor Moa, thank you for your support, for always encouraging my curiosity, while showing me how to navigate academic pursuits that have real-world impact. Richard, thank you for the difficult questions that made me and my work better. To my examiner – Devdatt – thank you for your feedback and support.

Lovisa, thank you for your unwavering perseverance and endless curiosity that was an invaluable source of optimism during our collaboration. Working with you has though me so much and has been a highlight of my PhD journey. Bastiaan, Kajsa, and Pasko – thank you for the stimulating discussions and continuously putting the work we do in a more human perspective – you provided a very needed context and knowledge. To Nicolas and Mehrdad – thank you for a much needed camaraderie during difficult submission deadlines and interesting discussions that kept me motivated the rest of the time. To all WASP HS students – thank you for making me feel understood and giving me a sense of belonging even across hundreds of kilometres.

To Dag, for teaching me so much about developing skills in others and myself. Marco and Annika, thank you for your example and sharing your invaluable experience and expertise, helping me develop my skills as a researcher. Many thanks to Clara, Fatima, Andrea, and the whole administrative team at Chalmers for their invaluable help in navigating the rules and procedures. My thanks to the DSAI division for their support and guidance and a special thank you to the Formal Methods unit for making me feel welcome at the beginning of my PhD journey.

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

This thesis presents work at the intersection of machine learning, in particular natural language processing (NLP), and social science, specifically political science. The abundance of textual social science data, the strong connection between language and culture, together with recent advancements in the field of computational methods for text and language modelling make this a promising synergy possibility. We explore the utility of these new methodologies as well as their shortcomings and caveats. The work is addressing two main research questions:

RQ1 How can neural network methods from the NLP field assist in conducting social science research?

RQ2 How robust and reliable are the results produced by these methods?

1.1 Language Processing and Social Science

The rapidly shifting landscape of how we model language in machines has resulted in both new avenues of research and new challenges in methodology development. The natural language processing field looks very different now compared to how it did at the beginning of this project, in 2020. The main paradigms were pre-training and fine-tuning, benchmarking, architecture development, recurrent networks (e.g., LSTMs), and smaller, specialized models. The introduction of the attention mechanism (Bahdanau, Cho and Bengio, 2015) and the transformer architecture (Vaswani et al., 2017), which facilitated a shift away from recurrence, adapting architectures to the available hardware and benefiting from parallelization, allowed for more efficient training and thus upscaling of models. The new paradigms became large language models (LLMs), zero-shot learning, task instructions, prompt-engineering, and large proprietary models. This thesis addresses some of the old paradigms, with works based on word embeddings and learning from smaller corpora, and the shift to LLMs with works focusing on LLM capabilities and inner workings.

During the same period we have also seen huge political shifts and a global pandemic, which induced a transformation in how information is produced

and shared to more digital spaces. Understanding and predicting political behaviour is as important as ever, leading to the desire to utilize new NLP tools for the study of the political landscape. With the emergence of new communication mediums and the ever-increasing complexity, the data, relevant to answering our political questions, has gone out of the realm that is possible for manual processing. Large scale digitalization efforts such as online access to governmental documents (for example, from the Swedish National Parliament (Riksdag)) are one source of such corpora. Methods that can handle large amounts of data through statistical analysis rather than manual one are needed.

Neural networks have emerged as dominant paradigms in modelling language due to their impressive capability to capture complex patterns. The primary focus for improving language model performance has been on increasing complexity through more sophisticated architectural components and increased model size. While these advancements yield significant improvements, they also raise crucial questions regarding robustness, interpretability, and ethical implications. As the intricate nature of language in political discourse demands both nuanced understanding and explainability, the consequences of utilizing these methodologies for the study of political behaviour are particularly impactful. In this work, we explore these considerations, analysing how current approaches align with the objectives of modelling political language and behaviour.

On the one hand, language models enable the processing of vast quantities of data, facilitating access to insights from diverse and, in many cases, previously unexplored sources. These models can also reveal new types of connections by identifying patterns and underlying semantic signals within text. Furthermore, language models offer novel affordances, such as generation – these models can create text based on input prompts, and simulation – they can generate synthetic samples to model linguistic interactions. In this work we investigate some of these affordances and how they can facilitate political science research.

On the other hand, these models have demonstrated robustness and reliability challenges. Their outputs can vary significantly with changes in hyperparameters, variations in dataset characteristics, and even due to intrinsic stochasticity during the training phase (e.g., random initialization). When the primary focus is on relations and entities of the real world, the model's internal representation and "world view" become critically important. Specifically, understanding what the model has internalized and how its representations align with real-world contexts is essential. This is because relying solely on performance metrics from held-out datasets is insufficient for evaluation, given the complex and nuanced nature of these tasks. This thesis aims to both measure and explore contributing factors to robustness in models as well as to bridge the gap to uncovering internal model representations through work in explainable AI (XAI).

Structure of thesis

The thesis is organized as follows: In Section 1.2 we introduce the distributional hypothesis, which is the theory of meaning that has allowed the development of new powerful models. We introduce the relevant models and applications in

Section 1.3. We then outline how the probabilistic nature of these approaches relates to their stability in Section 1.4. We end with a discussion of how the complexity of the models hinders evaluation and introduce explainability as an approach for alleviating these issues in Section 1.5.

In Chapter 2 we introduce a short summary of the included papers. Addressing **RQ1**, we investigate what political science insights we can get from textual corpora by utilizing NLP methods in:

Paper I Word embeddings on ideology and issues from Swedish parliamentarians' motions: a comparative approach

Paper II Class Explanations: the Role of Domain-Specific Content and Stop Words

Addressing **RQ2**, we investigating what influences model robustness and consistency in:

Paper III The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models

Paper IV Fact Recall, Heuristics or Pure Guesswork? Precise Interpretations of Language Models for Fact Completion

Finally we present a pilot study which has relevance to both **RQ1** by investigating simulation as a new affordance of LMs and **RQ2** by scrutinizing the stability and robustness of such methods:

Paper V Identifying Non-Replicable Social Science Studies with Language Models

In Chapter 3 we outline conclusions and future work.

1.2 Distributional Hypothesis and Meaning

The Distributional Hypothesis, which is a cornerstone for most approaches to modelling nowadays, can be summarized with Firth's well-known and highly-cited quote – *"you shall know a word by the company it keeps"* (Firth, 1957). The hypothesis is commonly attributed to both Harris (Harris, 1954) and Firth (Firth, 1957) and in essence states that we can represent words through the context in which they appear. This in practice means representing language through its statistical distribution – e.g., through collocations of words in a corpus. It is often claimed in NLP literature that what is captured with this approach is *meaning* (Bender and Koller, 2020), however it is worth discussing two caveats that relate to this issue.

Coming from a structuralist view Harris proposes language can be *"described in terms of a distributional structure* [...] *and* [...] *this description is complete without intrusion of other features such as history or meaning"* – that is, meaning is something that is derived from human experience and the structure of a language deviates in many respects from that external structure of meaning.

He poses however that those two aspects – the external notion of meaning and the distributional structure of a language – are highly interconnected and one important aspect is that of difference. Stating that *"difference of meaning correlates with difference of distribution"* he provides a very powerful connection between meaning and form.

Firth, on the other hand, comes from a more anthropological view of language and discusses at length what needs to be considered when we talk about "company" or – how do we define the relevant context from which we extract the distributional description. Whereas NLP approaches in this area tend to use the collocations of words within a specified window length in a corpus of text, Firth invokes the *context of situation* as a necessary parameter that needs to be accounted for. Firth provides a list of the three categories that comprise this idea of context: relevant features of participants, relevant objects, effect of the verbal action (Firth, 1957). He does not provide guidance for how to account for these in practice or how to determine relevance, but his ideas lay the groundwork for developments in the study of language in its social context.

One approach to extending the purely textual co-occurrence-based notion of representation and introduce a broader view of the social and physical context of language is comparative stratification. This is based on what Harris calls "sublanguages" and Firth calls "restricted languages" – broadly speaking, the idea that we can split up language use according to the setting within which it is used and observe different distributional characteristics which are representative of the meanings within that setting. An example of how this can be applied in practice is studying diachronic semantic shifts – that is, splitting up the corpus in time periods and extracting separate representations for each period. These can then be compared to track shifts through time. The corpora can however be stratified along any dimension of interest, which, as we discuss later, can be a powerful tool for studying political behaviour, where we can summarize a lot of domain knowledge into additional metadata along which we can produce meaningful splits of the data.

1.3 Neural Models of Language

In practice these theoretical ideas are usually implemented as measuring and modelling co-location patterns of words in textual corpora. These can be simple, corpus statistic methods, however, the dominant approach is complex neural network-based language models typically trained through proxy tasks such as predicting missing text. The affordances of models also range from using them for representation (as a proxy of meaning), to prediction (for example classification, sentiment analysis), to generation.

1.3.1 Feedforward Neural Networks

The simplest model used is a feedforward neural network – a network of interconnected nodes (representing simple activation functions), organized in

layers, with the output of each layer serving as input for the next one. An example of a widely used model within NLP based on this architecture is Word2Vec (Mikolov et al., 2013) – a neural network with a single hidden layer. It can be trained to predict either context (surrounding words) from a single word (skip-gram negative sampling – SGNS) or, alternatively, to predict a word given the surrounding context (continuous bag of words – CBOW). Once the network has been trained, we can use the internal weights of the hidden layer as static vector representations for the terms in the vocabulary.

1.3.2 Word Interactions and the Transformer

One limitation of the feed-forward neural network is the simplistic representation of language as a bag of words with no interaction. There have been several aproaches to model word interactions. These include recurrent neural networks (RNNs) and convolutional neural networks (CNNs). RNNs map each consecutive input to an output vector and a hidden vector, which is then used as input together with the next input in the sequence for mapping in the next step, while CNNs use kernels to aggregate information of several input positions such as the embedding vectors of neighbouring words. The dominant architecture currently in use, however, is the transformer (Vaswani et al., 2017), based on the attention mechanism (Bahdanau, Cho and Bengio, 2015).

A transformer has two components – 1) encoder, which maps text to numeric vectors (and can thus be used for contextual embeddings, as for example the trained transformer encoder model BERT (Devlin et al., 2019)); 2) decoder, which maps text (or a numeric vector) to text (and is thus used for generating text based on inputs, as for example the GPT models (Radford et al., 2018)). The main building block of the transformer is based on multi-head self attention (MHSA) and a feedforward neural network called multilayer perceptron (MLP). The model has several layers of processing, which produce internal representation vectors for each input position (token) x (we use X to denote the matrix containing all positional vector representations). The representation at layer l for token position i is calculated as (ignoring bias terms and layer normalization for brevity):

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l$$

where a_i^l is the output of MHSA component and m_i^l is the output of the MLP component. Attention is calculated based on the input representations from all token positions, while MLP is only calculated from the input representation at the current position.

The attention mechanism consists of constructing a query, key and value representations from the input representations by utilizing three projection matrices: W_Q^l, W_K^l, W_V^l.The output representation is then calculated as:

$$A^l = softmax(\frac{(X^{l-1}W_Q^l).(X^{l-1}W_K^l)}{\sqrt{d_k}} + M^l)(X^{l-1}W_V^l)$$

where d_k is the dimensionality of the key representations and M^l is a mask. For multi-head attention, this is done H times independently resulting in H

matrices: $\{A_0^l, A_1^l, ..., A_H^l\}$. These resulting representations are concatenated and projected by a matrix W_O^l to produce the final output matrix, the i^{th} row of this matrix is a_i^l – the attention representation at layer l for token position i.

The MLP component is applied only to the current token's internal representation and calculated with two projection matrices (W_{proj} and W_{fc}), a rectifying nonlinearity σ, and normalizing nonlinearity γ as:

$$m_i^l = W_{proj}^l \sigma(W_{fc}^l \gamma(a_i^l + x_i^{l-1}))$$

Since this architecture can be processed in parallel, large models, where multiple components are stacked together can be trained. These large models exhibit capabilities for inferences from few or even no examples, which has led to a new paradigm for human-model interaction, namely, instruction tuning. This is a fine-tuning step, during which a trained generative model is provided examples in order to learn how to follow task instructions (e.g., answer questions), allowing a more natural language interface with the model.

Finally, text retrieval has been proposed as a way to improve model performance. In essence this consists of an external textual databased that can be used to supply relevant context. This is typically done by embedding both the text entries of the database and the query then extracting the closest matches from the database (e.g., through cosine similarity). This can be done as part of the architecture, during the training process (e.g., Atlas (Izacard et al., 2023)) or can be an external component that can be used to augmented an already trained model by providing the relevant contexts in the instructions (Lewis et al., 2020).

1.3.3 Applications

These models have been found to capture different aspects of knowledge, political and social signals, which can facilitate the measurement, identification or analysis of different phenomena. These techniques have been used to survey political attention and conflict (Rheault and Cochrane, 2020; Rodman, 2020; Rodriguez and Spirling, 2022; Osnabrügge, Elliot and Morelli, 2021), investigate changes in word meaning over time (see e.g., Hamilton, Leskovec and Jurafsky, 2016, and Tahmasebi, 2018 for an example in Swedish), measure emotion in debates (Rheault et al., 2016), and as part of an analysis of ideological placement (Rheault and Cochrane, 2020).

Other applications explore using these models for generation rather than representation by using them to simulate human behaviour. For example, Argyle et al. (2023) show that conditioning LMs on socio-demographic information can generate political preferences similar to those of a representative human sample. Others have investigated whether behaviour and social preferences in economic games involving cooperation and coordination resemble those of humans (Akata et al., 2025; Aher, Arriaga and Kalai, 2023; Filippas, Horton and Manning, 2024), or whether models make human-like moral judgments (Dillion et al., 2025). In terms of cognitive ability, Strachan et al. (2024) compare LLMs with human performance on a battery of measures designed to measure different theories of mind.

In addition to complex, nebulous social signals, these models have also been shown to store, in some capacity, well-defined atomic factual information such as subject-relation-object tuple facts (Petroni et al., 2019; Dinan et al., 2019). Given the well-defined nature and the access to ground truth for comparison of factual information, studying how models behave in this setting and what affects their performance can give us a better understanding of their general strengths and weaknesses.

1.4 Probabilistic Models and Stability

A direct consequence of the distributional hypothesis is that the prevalent modelling paradigms are probabilistic in nature. As meaning can be captured through "company" or context, language can be viewed as a high-dimensional probability distribution of word co-occurrences. The goal of a model of language being to learn this distribution (typically from large corpora of data through some (proxy) learning task). Some old approaches are based on representing that probability directly as the co-occurrence frequency of words. More modern neural network-based approaches can loosely be seen as complex architectures with enough flexibility to represent complex functions, that are then trained on proxy tasks usually related to predicting missing text. As a result, the model can rank what is likely to co-occur (typically the next token) with the input text. However, even though these scores can and sometimes are normalized to represent a probability distribution over the model's vocabulary, it is not clear how aligned that probability distribution is to the high-dimensional probability distribution as which Harris claims language can be expressed. In fact there are several areas, where this distribution is shown to be misaligned with different concepts. For example, research on model calibration Jiang et al., 2021; Vasudevan, Sethy and Ghias, 2019 has shown that token probability does not align with accuracy and as such cannot be used as a good approximation of confidence.

It is clear that the learned distribution is, at best, some approximation of the idealized Harris distribution. This probabilistic approximation of language leads to issues with robustness and reliability. During training the distribution that is learned is sensitive to training data, training order, hyper-parameters, initial conditions, while during inference the predictions are sensitive to input variation and hyper-parameters.

1.5 Complexity and Knowledge

The neural network-based approach together with benchmark-driven development leads to many improvements at the expense of a deeper understanding of the underlying models and the nuances of language itself. For tasks that align closely with specific benchmarks, this may not pose significant issues, as model performance can be quantitatively evaluated against these datasets. However real-world tasks are rarely that well defined. Even less so, when we are interested in the representation that the model has learned and we don't

have a gold standard for the desired performance, making it impossible to assess model performance with relation to the "truth". Taken together with the issues of robustness discussed earlier, the need for guarantees on the model's knowledge in these cases is essential. The complex question "what does the language model know?" requires us to start with a brief discussion of what knowledge is and how can we access or measure it.

One widely accepted definitions of knowledge is "justified true belief", where *true* refers to accurate reflection of reality and *justified* refers to having support by reasoning or evidence. However, the Gettier problem presents counterexample scenarios where this definition falls short (Gettier, 1963), prompting extensions to include notions like "acts of intellectual virtue" (e.g., carefulness, thoroughness, and a desire for truth) as a means to circumvent such issues (Zagzebski, 2017) stating that a belief needs to not only be true, but be acquired through a process characterized by intellectual virtues. Recent findings (Fierro et al., 2024) suggest this definition to be the most widely accepted one both by computer science and philosophy scholars. This focus on justification, or even more stringently – on "acts of intellectual virtue" – however, requires access to the underlying causes or mechanisms for model predictions as means of providing justification and distinguishing true knowledge from, for example, guessing. In our pursuit of understanding knowledge within the context of language models, a central focus for this thesis is exploring explainability techniques to bridge the gap between opaque algorithmic functions and the human-centric notion of true knowledge.

There exist multiple approaches to explainability – from attribution and gradient-based approaches, to probes, to prompting. There have been several attempts to organise these methods into different taxonomies and categories (Danilevsky et al., 2020; Madsen, Reddy and Chandar, 2022). One framework, that we find particularly compelling and relevant to the current thesis, is based on work by Doshi-Velez and Kim, 2017; Madsen, Reddy and Chandar, 2022 and introduces two properties, central to explainability – "functionally-grounded" and "human-grounded" explanations. Technically speaking, neural networks are explainable in the sense that they are deterministic at inference time, so we can track through the calculations and see the causal connections between the activation of all neurons in all layers. The reason we need explainability methods is that these sequences are prohibitively large and complex for a human to process. Therefore, explainability is some abstraction from the original model. This gives rise to an obvious trade-off – explanations true to the models (*functionally-grounded*) vs explanations that are understandable by a human (*human-grounded*).

For human-grounded explanations, what we need to consider is the different capabilities in pattern recognition. A machine learning model has advanced statistical capabilities, allowing it to detect weak signals from large amounts of data that would be beyond a human's computational capabilities. However, they are also restricted by the data they have been trained on, by definition having no access to external information, whereas, a human brings their extensive world knowledge to any task they solve, which implies the reliance on and use of a much richer information and representation of the world. In other words,

models are better at detecting small distributional differences, whereas humans have domain and other types of external knowledge, which may lead them to "focus" on different pieces of information when making inferences. For example, a person could classify a text as right-leaning because they identify the message as calling for the lowering of taxes, while a model attaches importance to words such as "the", "and", etc., since those could have slightly different distributions between parties. While both can be valid patterns from a modelling perspective (and lead to better predictive accuracy) only the first type of features can be aligned with political background knowledge.

One framework that aims to address the question of model faithfulness as a requirement for functionally-grounded explanations is causality. This is the motivation and theoretical framework used for a group of explainability methods called "mechanistic interpretability". Works by Vig et al. (2020) and Geiger et al. (2021) argue that the complexity of the models can be represented as a directed network of states, which form a causal graph, and are thus subject to causal mediation analysis. In essence, this means, we can intervene on model states and measure causal effects of that state's contributions to model behaviour. A typical way in which this is done is by patching (replacing or nullifying) states.

A simple version of this concept is the LIME (Ribeiro, Singh and Guestrin, 2016) approach, which aims to nullify input tokens in order to measure the contribution of different words to the final model prediction. More sophisticated approaches aim to identify internal model components that facilitate the flow of information. Two examples we use in this thesis are causal tracing (CT) (Meng et al., 2022) and information flow (Geva et al., 2023).

LIME is used to approximate a black-box model locally in several steps: First, for a particular input instance, we create an augmented version by removing a random number of words from the input. This is repeated several times to obtain a new data set locally around the input. Second, the new data points are passed through the black-box model for inference to get their respective labels, essentially creating a new dataset locally around the input. Third, an explainable model such as regularized logistic regression or a decision tree is fit to this data (typically weighted by the distance between the new datapoints and the original input). Finally, we can use the explanations from this new model (e.g., the weights of the regression model) as attribution scores of the features for the original input instance that we aim to explain.

Causal tracing is a method to find model components which facilitate the information from a particular part of the input to increase the probability of a prediction. The method consist of three steps: 1) a clean run, where a model is run in inference and the model states are recorded; 2) a corrupted run, where the information from the input tokens of interest is cut off by adding noise to the embeddings; 3) a patched run, where model state values from run 1) are used to overwrite the values of run 2) and the shift in output probability is measured and used to indicate the importance of that state in transmitting the information. By replacing m_i^l between the clean and corrupted runs, we can

explore the role of the MLP component. Previous work finds this intervention to indicate a crucial state for factual predictions (Meng et al., 2022).

Information flow instead tests where information is propagated rather than stored. This is done by cutting paths between two positions. By using attention knockout, which severs attention heads between the last position before the output and a position of interest, the authors find crucial model states that propagate information relevant in fact completion settings. This is done by setting the relevant section of the M^l matrix to $-inf$. The drop in prediction probability is used to indicate that the severed connection is important for transmitting information from the position of interest to the output. In addition to this intervention, Geva et al. (2023) also investigate what information is propagated by using the final layer projection matrix (which maps the final layer representations to the vocabulary). This matrix is applied to earlier layer representations in order to cast them onto the vocabulary and thus get an estimation of what words of the vocabulary those representations score higher.

Chapter 2

Summary of Included Papers

2.1 Word embeddings on ideology and issues from Swedish parliamentarians' motions: a comparative approach

In Paper I, we investigate the utility of word embedding methods for capturing politically relevant signals in the Swedish parliament. Being a proportional representation system, the Swedish parliament is characterized by the parties' need to collaborate in coalitions to reach a majority (Bäck and Bergman, 2016), which may lead to less polarized views and thus smaller differences in the text distributions. Therefore patterns that can indicate agreement and disagreement of meaning and word use are of particular interest. We base our approach on previous works that successfully use embeddings as a tool to track semantic shifts through time (Rodman, 2020) and measure similarity of political parties (Goet, 2019). We additionally examine the effects of several design choices on the results and the stability of the embeddings.

Methodology. We base our work on a corpus of Swedish parliamentary motions that contain early signals of political direction from individual party members. To simplify the task of detecting differences in language use we focus on the two main parties representing the right-left political spectrum – Moderates and Social Democrats. We examine 10 terms covering topic indicated as important for voters (Fredén and Sikström, 2021) and compare how those are embedded differently between parties.

Based on previous work aimed at tracking overall word use change, we use static embeddings by training Word2Vec models. Due to limited data, we employ a transfer learning approach, leveraging "general language" learned as a starting point for adapting to the different strata of data. Since our task of learning embeddings is the same as the pre-training task, during adaptation we

opt for fine-tuning rather than feature extraction. We compare two pre-trained models – an external model trained on general Swedish text available from the Nordic Language Processing Library (NLPL) word embedding repository[1] and a model pre-trained on other Swedish parliamentary data.

Fine-tuning is done from both of these pre-trained models on each strata (party and time period combination). We then extract the 20 closest words in the embedding space to the term of interest based on cosine similarity. To estimate the stability of the resulting embeddings we perform a bootstrapping of the data by training 10 versions of the model for each strata and calculating the mean and standard deviation distances between vectors.

We evaluate the results by manual investigation of the top 20 closest words to the terms of interests. We additionally investigate the stability of results by looking at the number of words that are more than one standard deviation above the score of the twentieth word in the list (indicating that those appear in the list more reliably and are less likely to appear due to the stochastic nature of the model or small variations in the data).

Results. From manual investigation of the results, we find that some word associations correlate with expected party views. For example, in the crime dimension we see an association with tax crimes for the Social Democrats and with gang crime and assault for the Moderates. Additionally, in the solidarity dimension we observe association with peace and welfare for the Social Democrats and security and stability for the Moderates. These associations correspond well to the general tendencies of left-wing parties to center on social and economic welfare, whereas the right-wing focus more on security. When looking at the stability of the embeddings, we see the most salient difference is between terms (rather than between pre-training approaches or parties). Similarly to previous works (Wendlandt, Kummerfeld and Mihalcea, 2018; Borah, Barman and Awekar, 2021), we find a roughly logarithmic relationship between stability and frequency. However due to the large variation in this relationship, we suggest there might be other factors contributing to this effect more connected to the types of words – for example value laden versus policy terms.

Contributions. D. Saynova contributed to the design of the study, training the models, and aggregating and summarizing the results as well as the writing of the paper. A. Fredén and M. Johansson contributed to the design of the study, interpretation of the results, writing of the paper and supervision of the project.

[1] `http://vectors.nlpl.eu/repository/20/69.zip`

2.2 Class Explanations: the Role of Domain-Specific Content and Stop Words

In Paper II, we explore the state of current XAI methods and their utility for the social sciences. We identify the need for developing suitable class-level explanations and propose a novel approach that provides ranked feature lists for a binary text classifier that separates domain specific content words from stop words.

Methodology. We propose a four step algorithm for producing class-level explanations: First, we run an instance explainability method on a selected set of datapoints (in our application we use LIME). Second, we aggregate the instance-level explanation features into their respective lists for the two classes. Third, we propose two scoring approaches – one based on frequency normalization and one on principal component analysis of embeddings – to rank the feature lists along a dimension from domain specific to stop words. Finally, we propose the use of "keywords in context" (KWIC) to exemplify the texts in which those features appear and allow the examination of the validity of those patterns. We test this approach on a black box model (in our application a BERT classifier) trained to predict party from text. The corpus we use for the case-study contains debates transcripts from the Swedish Riksdag.

Results. Both our scoring functions result in domain specific words at the top of the lists and stop words at the bottom, with the normalization approach resulting in mainly function words at the bottom of the lists. We further see that the top words refer to taxes and employment which reflects the studied texts and the left/right dimension in Sweden. Through deeper analysis with KWIC of the term "labor market policy" (identified as important for Social Democrats) we show how these features can be validated with domain knowledge. Finally, based on a small sample of datapoints, we find that the model performs better for texts that have predominantly domain specific content word features as explanations.

Contributions. D. Saynova contributed to the design of the study and the proposed novel XAI method, implementation of the methods as well as the writing of the paper. B. Bruinsma contributed to the design of the study and the XAI method, providing the political science context and framing of the studied materials and writing of the paper, M. Johansson and R. Johansson contributed to the design of the study and the XAI method as well as writing of the paper and provided supervision for the project.

2.3 The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models

In Paper III, we further investigate the aspect of robustness. We are specifically interested in exploring consistency of language models in fact completion scenarios. Models exhibit a tendency to produce different answers to the same semantic query when it is rephrased. For example, a model might predict both "Anne Redpath passed away in Edinburgh" and "Anne Redpath's life ended in London.". We focus on this constrained setting of subject-relation-object queries and we measure consistency under two model paradigms – upscaling (increasing model size) and retrieval augmentation. We also explore the characteristics of the prompt that influence consistency.

Methodology. We use the ParaRel dataset (Elazar et al., 2021), which consists of a collection of subject-relation-object tuples (e.g., "Anne Redpath"-"died in"-"Edinburgh") and a set of phrasings for expressing the subject and relation in a way that the object should be generated by the model (e.g., "Anne Redpath passed away in". We measure consistency as the number of rephrased pairs that produce the same answer normalized by all possible pairs (this depends on how many rephrased alternatives are available).

We test the consistency of two types of models: generative transformer-based LMs (in particular, we test LLaMA with sizes 7B, 13B, 33B and 65B) and retrieval-augmented models (we use Atlas in both base and large version with a Wikipedia retrieval corpus). We also explore characteristics in the ParaRel construction, focusing on surface level signal, that may influence prediction and measure consistency differences. We focus on three groups of phenomena: semantic overlap in answers, unidiomatic language (on both template and object level), and subject-object similarity. Semantic overlap affects several relations with more ambiguity in the possible responses – e.g., when a model is allowed to respond both "Edinburgh" and "Scotland". Unidiomatic language refers to some disruption in the flow of the query – e.g., when a model is expected to produce the sentence "Anne Redpath died at Edinburgh". Finally, overlap refers to cases where the names of the subject and objects share some substrings – e.g., "Nokia N9 was produced by Nokia".

Results. We find that both retrieval augmentation and upscaling lead to a considerable improvement of consistency. Upscaling shows diminishing returns with larger sizes as the performance plateaus. Additionally, retrieval augmentation is parametrically more efficient, with the Atlas base model (330M parameters) outperforming the largest, 65B, LLaMA model. We find that all query-related groups we investigate have an effect on consistency across all models, with performance dropping up to 40% in some cases.

Contributions. L. Hagström implemented and evaluated the Atlas models. She also contributed to the deeper investigations into causes of (in)consistency.

She also made major contributions to the writing of the paper. D. Saynova helped evaluate the models. She also contributed to the deeper investigations into causes of (in)consistency. She also made major contributions to the writing of the paper. T. Norlund implemented and evaluated the Llama models. He also consulted on the writing of the paper. M. Johansson provided supervision on the work and writing of the paper. R. Johansson provided supervision on the work and writing of the paper. He also helped write the paper.

2.4 Fact Recall, Heuristics or Pure Guesswork? Precise Interpretations of Language Models for Fact Completion

In Paper IV, we continue to build on concepts of knowledge and explainability. In this work, we revisit the problem of fact completion in language models and use tools from mechanistic interpretability to study the inner workings of language models. We expand on previous work, which assumes accurate prediction to be indicative of knowledge, and propose a more nuances taxonomy of prediction scenarios. We experiment with two mechanistic interpretability approaches to show the distinct model processing involved in these scenarios.

Methodology. Previous work on understanding how models use internal components to store and predict facts overrelies on accuracy as the sole indicator of knowledge and memorization. Our previous investigations in Paper III, as well as other work in the area, however, suggest that models that exhibit accurate predictions fail to behave in a way consistent with possessing knowledge. We propose three criteria that split model predictions into four scenarios: exact fact recall, heuristics recall, guesswork, and generic language modelling. Based on our taxonomy we build testing data for Llama 2 (7B and 13B) and GPT2-XL, containing examples of each scenario. We apply the methods of causal tracing (Meng et al., 2022) and information flow (Geva et al., 2023) to our subsets and compare the internal workings of the models.

Results. Our experiments with causal tracing show that previous conclusions of fact recall being facilitated by MLP layers at the subject token operating as a key-value store are consistent with the patterns we observe for the exact fact recall scenario. Furthermore, we observe patterns in general language modelling to not indicate the same components involvement, indicating these conclusions are indeed indicative of fact recall and not simply an artefact of the intervention. However, we find that heuristics and guesswork, while possible to produce accurate predictions, do not indicate key-value stores as the only and most influential internal mechanism. Our experiments with information flow further confirm the difference in internal components involved in the processing and prediction of exact fact recall, heuristics and guesswork.

Contributions. D. Saynova contributed to the identification method of prediction scenarios and to the causal tracing evaluations. She also made major contributions to the writing of the paper. L. Hagström contributed to the identification method of prediction scenarios and to the causal tracing evaluations. She also implemented the information flow analysis and made major contributions to the writing of the paper. M. Johansson provided supervision on the work and writing of the paper. R. Johansson provided supervision on the work and writing of the paper. He also contributed to the writing of the paper. M. Kuhlmann provided supervision on the work and writing of the paper. He also contributed to the writing of the paper.

2.5 Identifying Non-Replicable Social Science Studies with Language Models

In Paper V, we investigate if models can encode a more complex type of knowledge – namely, social views and attitudes. Motivated by the replication crisis in social science and evidence of language models exhibiting certain alignments with human preference, we explore LMs' suitability to indicate the replicability of a social science experimental study.

Methodology. We base our exploration on 14 text-based social science studies from the ManyLabs 2 replication project and use the human replication result as a true label of a study's replicability. We generate synthetic samples by querying LMs with the studies' texts and calculate new effect sizes and significance levels based on these synthetic samples. By comparing the LM results with human results, we measure the success of a LM in indicating replicability. We test four instruction-tuned LMs – GPT 4o and three open-source models: Llama 3 8B, Qwen2 7B, and Mistral 7B.

We also run temperature sensitivity experiments. One challenge with using LMs in this setting is that they tend to produce responses with low variance (Park, Schoenegger and Zhu, 2024). As the standard deviation is used to calculate effect sizes, this can lead to biased and inflated results. We thus run the experiment with four temperature settings and measure the change in effect size with temperature.

Results. We find the larger proprietary model and the smaller open-source ones perform similarly in terms of F1 scores when compared to the true (human) replication label. They achieve F1 of up to 0.77. Models also indicate higher precision and lower recall, meaning they are unlikely to replicate a non-replicable study and struggle to replicate even the replicable ones. Our temperature results confirm our hypothesis that a reduced variance can lead to an artificial increase in effect size, as we notice a general trend of larger effect sizes for lower temperature settings. Additionally, at the two lowest temperature settings we observe 12 cases when effect sizes exceed a magnitude of 4.0, with largest effect size of 21.6.

Contributions. D. Saynova contributed to the design of the study and ran the LM experiments, she also contributed to the writing of the paper and provided the technical background for the work. K. Hansson selected the studies and provided the social science motivation and context. She lead the interpretation of the studies and the effect size results. She also contributed to the writing of the paper. B. Bruinsma helped with the interpretation of the studies and the results, he produced and formatted the aggregated LM results. He also contributed to the writing of the paper and replicated the results of the previous studies to ensure consistent comparisons. M. Johansson and A. Fredén contributed to the design and motivation of the study, they provided supervision for the project.

Chapter 3

Discussion and Future Work

The main questions this work aims to address are: *How can NLP methods contribute to social science research?* and *How robust are these methods and the corresponding insights?*

Through the work carried out in this thesis, we find evidence that computational methods can support the study of political behaviour. We find computational models of language are capable of capturing important social signals, such as political leanings, policy and values. In particular we provide evidence of the effects of different training approaches for word embeddings. We show that off-the-shelf models can be used to detect policy and value differences between parties even in proportional representation systems such as Sweden. Our work contributes to growing empirical evidence, providing political insights from previously under-explored sources. We also provide a new method for assessing the alignment between human and model decision making, by developing an algorithm for extracting class-level explanations for model prediction, which is an under-explored area within explainability research. We look at new approaches to social science research facilitated by new model affordances and provide initial insights into how models can be used for generating synthetic data. We add to a growing body of work in this area by providing empirical evidence of the suitability of open-source and proprietary model-generated data for detecting replicable and non-replicable studies. One main theme throughout the work presented here is the question of reliability and robustness. When using these methods for political and social science research, we are interested in measuring real-world effects, associations, beliefs, and behaviours. The strong connection between language and culture means these phenomena can leave traces in textual data and from there be captured through modelling. Therefore questions of evaluating how well the learned connections reflect the real world and what effect that has on downstream applications are crucial. One imperative aspect of the NLP pipeline for social science remains the need for domain expert knowledge, continuous validity assessments, and improvements in model understanding.

We look at model robustness through a series of works studying a narrowly defined atomic version of factual information. We find that models still struggle even in those limited settings and show that two of the main model development approaches (upscaling and retrieval augmentation), while supporting model consistency, are insufficient to resolve the issue. Furthermore, we find several prompt-related signals which have a large effect on consistency. Finally, we contribute to the understanding of the internal mechanisms of language models for fact storage and retrieval, by breaking away from the previously used definitions of knowledge as accuracy and provide tools and analysis of the nuanced scenarios of language model behaviour. We conclude that there is still a need to develop theoretical frameworks, methodologies and testing protocols in order to bridge the gaps between model performance and behaviour and human needs and assumptions. The rise of proprietary models making this need even more pronounced.

3.1 Future Directions

In the final pilot study presented in this work, we investigate the utility of LLMs for simulation and synthetic data production. Due to the rapid pace and scale at which models can produce data, as well as the uncertainties still remaining in the field, we need to ensure (or at least quantify) the validity of these samples. We find that in this new, rapidly developing field, there is still a lack of guarantees of validity in relation to the real world. One way in which this can be assessed is expanding the empirical evidence by comparing to larger collections of human data. Other avenues of research may focus on increasing our toolset of explainability approaches such that we have a more direct access to model representations and "reasoning".

Finally, our work in explainability and model robustness suggests that there is a need to develop and strengthen our theoretical frameworks for understanding language model behaviour. Questions such as "what does a language model know?", "why does a model behave like this?", "what are the implications of certain model behaviours?", while crucial, require substantial theoretical work, as relying on frameworks used for human subjects (such as asking) or statistical models (such as testing on a held out set) are flawed and insufficient.

Bibliography

Aher, Gati, Rosa I. Arriaga and Adam Tauman Kalai (2023). "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies". In: *International Conference on Machine Learning*. PMLR. JMLR.org, pp. 337–371. DOI: 10.5555/3618408.3618425 (cit. on p. 8).

Akata, Elif et al. (2025). "Playing repeated games with large language models". In: *Nature Human Behaviour*, pp. 1–11 (cit. on p. 8).

Argyle, Lisa P. et al. (2023). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3, pp. 337–351. DOI: 10.1017/pan.2023.2 (cit. on p. 8).

Bäck, Hanna and Torbjörn Bergman (2016). "The Parties in Government Formation". In: *The Oxford Handbook of Swedish Politics*. Ed. by Jon Pierre. Oxford: Oxford University Press, pp. 206–226 (cit. on p. 13).

Bahdanau, Dzmitry, Kyung Hyun Cho and Yoshua Bengio (Jan. 2015). ""Neural machine translation by jointly learning to align and translate"". English (US). In: 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015 (cit. on pp. 3, 7).

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 (cit. on p. 5).

Borah, Angana, Manash Pratim Barman and Amit Awekar (2021). "Are Word Embedding Methods Stable and Should We Care About It?" In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. New York, NY: Association for Computing Machinery, pp. 45–55. DOI: 10.1145/3465336.3475098 (cit. on p. 14).

Danilevsky, Marina et al. (2020). "A Survey of the State of Explainable AI for Natural Language Processing". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. ACL, pp. 447–459 (cit. on p. 10).

Devlin, Jacob et al. (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)*, pp. 4171–4186 (cit. on p. 7).

Dillion, Danica et al. (2025). "AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise". In: *Scientific Reports* 15.4084. DOI: 10.1038/s41598-025-86510-0 (cit. on p. 8).

Doshi-Velez, Finale and Been Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. DOI: 10.48550/ARXIV.1702.08608 (cit. on p. 10).

Elazar, Yanai et al. (2021). "Measuring and Improving Consistency in Pretrained Language Models". In: *Transactions of the Association for Computational Linguistics* 9. Ed. by Brian Roark and Ani Nenkova, pp. 1012–1031. DOI: 10.1162/tacl_a_00410. URL: https://aclanthology.org/2021.tacl-1.60/ (cit. on p. 16).

Fierro, Constanza et al. (Nov. 2024). "Defining Knowledge: Bridging Epistemology and Large Language Models". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 16096–16111. DOI: 10.18653/v1/2024.emnlp-main.900. URL: https://aclanthology.org/2024.emnlp-main.900/ (cit. on p. 10).

Filippas, Apostolos, John J. Horton and Benjamin S. Manning (2024). "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" In: *Proceedings of the 25th ACM Conference on Economics and Computation*. EC '24. ACM, pp. 614–615. DOI: 10.1145/3670865.3673513 (cit. on p. 8).

Firth, John (1957). "A synopsis of linguistic theory, 1930-1955". In: *Studies in linguistic analysis*, pp. 10–32 (cit. on pp. 5, 6).

Fredén, Annika and Sverker Sikström (2021). "Voters' Sympathies and Antipathies Studied by Quantitative Text Analysis: Evidence from a two-wave panel experiment in Sweden during covid-19". In: *Annual Midwest Political Science Association Conference*. MPSA, Chicago (cit. on p. 13).

Geiger, Atticus et al. (2021). "Causal Abstractions of Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 9574–9586 (cit. on p. 11).

Gettier, Edmund L (1963). "Is justified true belief knowledge?" In: *analysis* 23.6, pp. 121–123 (cit. on p. 10).

Geva, Mor et al. (Dec. 2023). "Dissecting Recall of Factual Associations in Auto-Regressive 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. 12216–12235. DOI: 10.18653/v1/2023.emnlp-main.751. URL: https://aclanthology.org/2023.emnlp-main.751/ (cit. on pp. 11, 12, 18).

Goet, Niels D. (2019). "Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015". In: *Political Analysis* 27.4, 518–539. DOI: 10.1017/pan.2019.2 (cit. on p. 13).

Hamilton, William L., Jure Leskovec and Dan Jurafsky (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, pp. 1489–1501. DOI: 10.18653/v1/P16-1141 (cit. on p. 8).

Harris, Zellig S (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162 (cit. on p. 5).

Izacard, Gautier et al. (2023). "Atlas: Few-shot learning with retrieval augmented language models". In: *Journal of Machine Learning Research* 24.251, pp. 1–43 (cit. on p. 8).

Jiang, Zhengbao et al. (2021). "How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering". In: *Transactions of the Association for Computational Linguistics* 9. Ed. by Brian Roark and Ani Nenkova, pp. 962–977. DOI: 10.1162/tacl_a_00407. URL: https://aclanthology.org/2021.tacl-1.57/ (cit. on p. 9).

Lewis, Patrick et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf (cit. on p. 8).

Madsen, Andreas, Siva Reddy and Sarath Chandar (2022). "Post-Hoc Interpretability for Neural NLP: A Survey". In: *ACM Computing Surveys* 55.8, pp. 1–42. DOI: 10.1145/3546577 (cit. on p. 10).

Meng, Kevin et al. (2022). "Locating and Editing Factual Associations in GPT". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 17359–17372. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf (cit. on pp. 11, 12, 18).

Mikolov, Tomás et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013*. arXiv: 1301.3781 (cit. on p. 7).

Osnabrügge, Moritz, Ash Elliot and Massimo Morelli (2021). "Cross-Domain Topic Classification for Political Texts". In: *Political Analysis*. DOI: 10.1017/pan.2021.37 (cit. on p. 8).

Park, Peter S., Philipp Schoenegger and Chongyang Zhu (2024). "Diminished Diversity-Of-Thought in a Standard Large Language Model". In: *Behavior Research Methods* 56, pp. 5754–5770. DOI: 10.3758/s13428-023-02307-x (cit. on p. 19).

Radford, Alec et al. (2018). "Improving language understanding by generative pre-training". In: (cit. on p. 7).

Rheault, Ludovic and Christopher Cochrane (2020). "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora". In: *Political Analysis* 28.1, pp. 112–133. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2019.26 (cit. on p. 8).

Rheault, Ludovic et al. (2016). "Measuring emotion in parliamentary debates with automated textual analysis". In: *PloS one* 11.12, e0168843 (cit. on p. 8).

Ribeiro, Marco, Sameer Singh and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. ACL, pp. 97–101. DOI: 10.18653/v1/N16-3020 (cit. on p. 11).

Rodman, Emma (2020). "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors". In: *Political Analysis* 28.1, pp. 87–111. DOI: 10.1017/pan.2019.23 (cit. on pp. 8, 13).

Rodriguez, Pedro and Arthur Spirling (2022). "Word Embeddings: What works, what doesn't, and how to tell the difference for applied research". In: *Journal of Politics*. DOI: https://doi.org/10.1086/715162 (cit. on p. 8).

Strachan, James W. A. et al. (2024). "Testing Theory of Mind in Large Language Models and Humans". In: *Nature Human Behaviour* 8.7, pp. 1285–1295. DOI: 10.1038/s41562-024-01882-z (cit. on p. 8).

Tahmasebi, Nina (2018). "A Study on Word2Vec on a Historical Swedish Newspaper Corpus". In: *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, DHN 2018, Helsinki, Finland, March 7-9, 2018.* Pp. 25–37 (cit. on p. 8).

Vasudevan, Vishal Thanvantri, Abhinav Sethy and Alireza Roshan Ghias (2019). "Towards Better Confidence Estimation for Neural Models". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7335–7339. DOI: 10.1109/ICASSP.2019.8683359 (cit. on p. 9).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. (cit. on pp. 3, 7).

Vig, Jesse et al. (2020). "Investigating Gender Bias in Language Models Using Causal Mediation Analysis". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 12388–12401. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf (cit. on p. 11).

Wendlandt, Laura, Jonathan K. Kummerfeld and Rada Mihalcea (2018). "Factors Influencing the Surprising Instability of Word Embeddings". 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, LA: Association for Computational Linguistics, pp. 2092–2102. DOI: 10.18653/v1/N18-1190 (cit. on p. 14).

Zagzebski, Linda (2017). "What is knowledge?" In: *The Blackwell guide to epistemology*, pp. 92–116 (cit. on p. 10).