Thesis for The Degree of Doctor of Philosophy

Representations, Retrieval, and Evaluation in Knowledge-Intensive Natural Language Processing

Lovisa Hagström

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

Representations, Retrieval, and Evaluation in Knowledge-Intensive Natural Language Processing

Lovisa Hagström

ISBN 978-91-8103-250-5
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5708.
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.

Representations, Retrieval, and Evaluation in Knowledge-Intensive Natural Language Processing

Lovisa Hagström

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

Abstract

Several major advancements have recently been made within the field of Natural Language Processing (NLP). Nowadays, NLP systems based on language models (LMs) are readily available to the public in the form of chatbots, code assistants, writing assistants, etc. Any task that can be described in text can be, and is, addressed by NLP systems, covering the expected tasks as well as less expected tasks. While these advancements have highlighted many strengths of NLP systems, they have also highlighted weaknesses of NLP systems, hindering their use in certain scenarios. For example, modern NLP systems are neither reliable nor interpretable, limiting their usefulness for e.g. knowledge-intensive or high-risk tasks. In this thesis, we focus on the application of NLP systems to knowledge-intensive situations. We consider how methods leveraging different types of representations of information, such as the parametric memory of a model trained on multimodal information or retrieval-augmented generation (RAG), can be used to improve the systems. We find that RAG can be used to improve the stability of NLP systems for knowledge-intensive tasks, and bigger LMs generally are more efficient in leveraging the external information in RAG. We also develop datasets and methods to allow for more comprehensive and precise evaluations of NLP systems in knowledge-intensive situations. We find that insights gained from synthesised evaluation datasets are not guaranteed to transfer to real-world scenarios and that evaluation results are sensitive to how the knowledge under consideration interacts with the parametric memory of the LM. Taken together, the work included in this thesis improves our understanding of NLP systems for knowledge-intensive situations and highlights the important role of representations of information as well as realistic benchmarks for NLP.

Keywords

Natural Language Processing, Knowledge-intensive Tasks, Retrieval-Augmented Generation, Language Models, Context Utilisation, Mechanistic Interpretability, Vision-and-Language Models, Evaluation

List of publications

Appended publications

This thesis is based on the following publications:

[**Paper 1**] T. Norlund*, **L. Hagström***, R. Johansson, *Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?*
Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (Nov 2021), 149-162.

[**Paper 2**] **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*
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Dec, 2023), 5457–5476.

[**Paper 3**] **L. Hagström**, S. V. Marjanović, H. Yu, A. Arora, C. Lioma, M. Maistro, P. Atanasova, I. Augenstein, *A Reality Check on Context Utilisation for Retrieval-Augmented Generation*
Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Jul, 2025), 19691–19730.

[**Paper 4**] D. Saynova*, **L. Hagström***, M. Johansson, R. Johansson, M. Kuhlmann, *Fact Recall, Heuristics or Pure Guesswork? Precise Interpretations of Language Models for Fact Completion*
Findings of the Association for Computational Linguistics: ACL 2025 (Jul, 2025), 18322–18349.

[**Paper 5**] **L. Hagström***, Y. Kim*, H. Yu, S. Lee, R. Johansson, H. Cho, I. Augenstein, *CUB: Benchmarking Context Utilisation Techniques for Language Models*
Under review.

[*] 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**] **L. Hagström**, R. Johansson, *Knowledge Distillation for Swedish NER models: A Search for Performance and Efficiency*
Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa) (May, 2021), 124-134.

[**b**] **L. Hagström**, R. Johansson, *What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge*
Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (May, 2022), 252-261.

[**c**] **L. Hagström**, T. Norlund, R. Johansson, *Can We Use Small Models to Investigate Multimodal Fusion Methods?*
Proceedings of the 2022 CLASP Conference on (Dis)embodiment (Sep, 2022), 45-50.

[**d**] **L. Hagström**, R. Johansson, *How to Adapt Pre-trained Vision-and-Language Models to a Text-only Input?*
Proceedings of the 29th International Conference on Computational Linguistics (Oct, 2022), 5582-5596.

Acknowledgment

I would first like to thank my family. Mattias, thank you for your endless love, support and advice. You have spent so many hours helping me polish papers, presentations and posters, even when we both probably would have preferred doing something else. Sofia, thank you for always being there and for your patience when I needed to work more than I promised. Dad, thank you for always being interested and invested in my work. I remember when you worried about whether I would be out of a job when ChatGPT was released. Mum, thank you for always being ready to lend an ear and for reminding me of what is important. Anna, thank you for your steadfast support. Thank you Farmor, you never had the same opportunities as me in life but was always so happy for me and all of your grandchildren. Mormor, thank you for your love, mainly expressed in wonderful food and textile crafts. Thank you Farfar, you never got to know that your grandchild would achieve a doctor of philosophy, while it would not have mattered as you always was so proud of and happy for all of your grandchildren.

I would also like to thank my colleagues who helped make these five years of PhD studies so much more fun and rewarding. Richard, thank you for your unwavering support and encouragement throughout weekends, summer holidays and tough times. Denitsa, thank you for the many project collaborations – working with you always felt so easy. Marco, thank you for always being helpful and supportive. Isabelle, thank you for believing in me and for gifting me with so many research opportunities. To Nicolas, Mehrdad, Tobias and Denitsa who were a part of the NLP lab, thank you for contributing to a conducive and ambitious work environment. To the CopeNLU lab I visited in Copenhagen, thank you for taking me in with so much warmth and for enduring my project leader efforts. Thank you Lena, my colleague in crime, who made the years of PhD studies so much more enjoyable. Thank you to my office mates, Adam, Anton, Lena and Newton, who started at the same time as me and have remained reliable friends throughout covid and life. To all of my colleagues at the DSAI division, thank you for contributing to a work environment in which we all respect and value each other. To my colleagues across universities and national borders, thank you for all of the memorable experiences throughout these five years. Thank you Graham for making the administrative parts of PhD studies seem so simple. Lastly, thank you to the research school at Chalmers for your support and compassion.

Contents

II Appended Papers 55

Part I

Summary

Chapter 1

Introduction

The field of natural language processing (NLP) is currently in its *deep learning* era, for which *artificial neural networks* (ANNs) are used to model language. At the start of this era, in the 2010s, one of the main goals was to generate coherent text (Wang et al., 2018). For example, when OpenAI presented their GPT-2 model[1] in 2019, they used the example of a generated news article on talking unicorns to showcase the impressive abilities of the model (Radford et al., 2019). In the recent six years, NLP systems have advanced well beyond the problem of generating coherent text. The research frontline now spans a plethora of more difficult and unsolved tasks, many of which are so called *knowledge-intensive tasks*, i.e. tasks humans cannot be expected to solve based on memory alone, for which access to some form of knowledge source is crucial (Petroni et al., 2021). Examples of such tasks are question-answering and fact-checking tasks.

NLP systems suffer from issues related to *hallucinations*, i.e. the propensity to generate text that appears coherent but contradicts factual knowledge or system input, and *unpredictable instability*, i.e. seemingly insignificant changes in input may cause critical changes in performance (Maynez et al., 2020; Elazar et al., 2021). These issues and our inability to interpret the blackbox systems ultimately make most NLP systems too unreliable for safe use in knowledge-intensive tasks. For these tasks, the user needs to be able to trust the system, as they generally are unable to verify the answer themselves, and the benefit of using the NLP system relies on not having to verify the answer. This task is different from e.g. the task of rephrasing text for which the user easily can verify the generated text themselves.

To address these issues, new system designs have been proposed. The systems combine ANNs with external sources of information, such as Wikipedia, the web or knowledge graphs (Lewis et al., 2020; Shuster et al., 2021; Gao et al., 2024). These systems are typically referred to as *Retrieval-Augmented Generation* (RAG) or *Retrieval-Augmented Language Models* (RALMs). Recent results have shown these system designs to mitigate issues with hallucinations and instability in knowledge-intensive situations.

[1] A predecessor to ChatGPT.

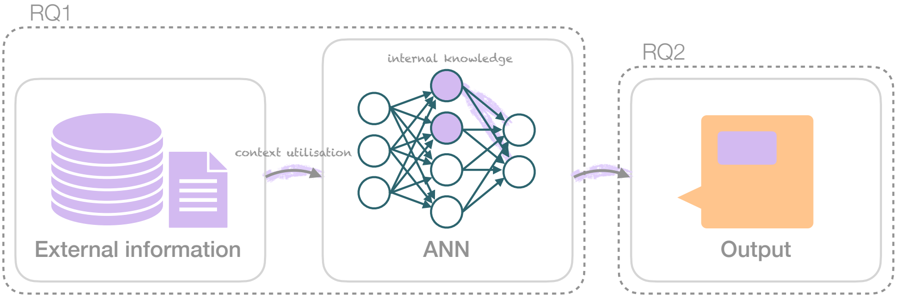


Figure 1.1: An illustration of NLP systems for knowledge-intensive situations, potentially combining external information from e.g. the web with ANNs to generate an output. Potential sources and transfer of information are highlighted in purple. The research questions discussed in this thesis are also depicted with the dashed boxes.

Recent results have also found ANNs capable of storing information in their parameters (Petroni et al., 2019). However, this storage is seemingly too limited and unstable compared to alternatives that leverage external sources of information.

These recent developments raise interesting questions related to how to represent and leverage information[2] in NLP systems for knowledge-intensive tasks, see Figure 1.1. For effective system designs, should the information come from trained ANN parameters, or external representations of information, and does it depend on the type of information? Furthermore, for trained ANNs, how is information stored? For the evaluation of NLP systems, how can we know the source of the outputted information? And how is information transferred in NLP systems – is external information, when provided, always incorporated in the ANN output? Much of the work included in this thesis focuses on these questions, summarised by the following exploratory research question.

> *RQ1: How is information transferred in NLP systems and how should it be transferred for effective system designs for knowledge-intensive tasks?*
> Addressed in papers 1, 2, 4 and 5.

We also need sound and relevant evaluation methods to help us ascertain whether novel NLP systems improve key traits for knowledge-intensive tasks. Due to the blackbox nature of modern NLP systems, their performance can only be measured via empirical approaches rooted in evaluation data that has been designed to elicit and test the trait under consideration. Failures in designing appropriate evaluation data may result in misleading conclusions that do not generalise to the areas of interest (McCoy et al., 2019; Zellers et al., 2019). The development of sound evaluation data comes with many challenges; manual annotation of data is the most appropriate approach, but comes with high

[2] In this thesis, we mainly consider representations of information with stored or learned information about the world, see Chapter 3.

costs in money and time. Something that is solved by synthesising data, but with the risk of inducing unwanted artefacts that interfere with the evaluation, resulting in findings that do not generalise to real-world scenarios. This leads us to the second and final exploratory research question investigated by the work included in this thesis.

> *RQ2: How should we evaluate NLP systems for knowledge-intensive situations?*
> Addressed in papers 1, 3, 4 and 5.

The content of this thesis starts by introducing models for natural language processing (§2). We then move on to consider how representations of information are used to support NLP systems (§3) and how NLP systems are evaluated in knowledge-intensive situations (§4). Finally, the work included in this thesis is summarised (§5) and we consider final conclusions together with reflections on future work (§6).

Chapter 2

Models for natural language processing

Given some text input like “Q: What is the colour of the sky? A:...” or “This movie was awesome!”, models for NLP are used to infer adequate outputs like “Blue” or “Sentiment: Positive”. ANNs are widely used in NLP by virtue of their adaptability and capacity to learn from unsupervised training on data (Hornik et al., 1989), making them especially suitable for language processing.

A type of neural network that has found wide applicability in the field of NLP is the *Transformer* (§2.1). Most modern NLP systems are based on the Transformer. Depending on application area, different categories of NLP models are used, corresponding to suitable modifications to the Transformer network. The works included in this thesis have mainly focused on models for language representation (§2.2), language models (§2.3), and vision-and-language models (§2.4), all modelled using the Transformer.

2.1 Transformer

The Transformer network is used for most modern NLP models (Vaswani et al., 2017). This deep network utilises stacked *attention* layers to model dependencies between words in a sequence and has proven to be very performant for NLP (Bahdanau et al., 2015). This network setup works well also for longer sequences where the model has to take long-distance word-to-word relationships into account. Compared to the previous state-of-the-art NLP models based on recurrence and convolutions, the Transformer architecture largely avoids sequential computing. Thanks to its superior modelling capacity and parallelisability, the Transformer is the current state-of-the-art network for language processing.

The Transformer network was originally developed for language translation. Since translation is a sequence-to-sequence task, the original Transformer architecture consisted of two networks, an *encoder* network to encode the input to be translated and a *decoder* network that generates the translation based on the encoded input and preceding output. Each of these network parts can be, and have been, used separately in modern NLP models. The Transformer encoder lends itself especially useful for representation learning, while many design aspects of the Transformer decoder are useful for autoregressive language modelling, for which we wish to generate continuations based on preceding values of a provided sequence. The only remaining sequential aspect of this model is the generation by the decoder, meaning that all other computations can be parallelised for faster training.

Both the encoder and decoder of the Transformer network build on stacks of respective identical layers. The layer for the encoder consists of multi-head attention and a fully connected feed-forward network components. The layer for the Transformer decoder is similar to the encoder layer, while it contains additional attention over the encoder output and masks the attention over the decoder input to prevent information leakage from the tokens to be predicted, i.e. *causal attention*. This stacked setup allows for easy re-scaling of the Transformer, since one can simply change the number of layers in the stacks.

In the subsequent sections we explore how the Transformer is used for three different categories of NLP models (language representations, language models and vision-and-language models). While we here distinguish between the different categories, it can generally be assumed that many of the insights gained from the study of one model category transfer to the other, by virtue of all models being based on the Transformer. For example, insights related to the text processing performance and behaviour of models for language representation are likely to transfer to language models.

2.2 Language representations

An NLP model can be used to generate a vector, a representation, of the text input that then can be used instead of the text for any text related task, such as sentiment classification or categorisation (Peters et al., 2018; Devlin et al., 2019). Typically, a Transformer *encoder* is used for representation modelling.

A model frequently used for language representation is the *Bidirectional Encoder Representations from Transformers* (BERT) model. It had a large impact on modern NLP research after it was developed by Devlin et al. (2019), and showed a promising path forward for NLP. BERT is a language representation model that has been trained to generate contextualised token representations in a bidirectional fashion, also considering the words after the word of interest in a sequence. The BERT model comes in two sizes, BERT-base and BERT-large, modelled by a Transformer encoder with stacks of 12 or 24 layers respectively. The encoder generates representations for input tokens that can be used by a smaller network to solve some downstream task, as illustrated in Figure 2.1. The assumption is that if the encoder is sufficiently trained, it should be able to generate language representations that are useful for generic language tasks, as in transfer learning.

The BERT model is trained in two steps. The first step is a pre-training phase in which the model is tasked with Masked Language Modelling (MLM),

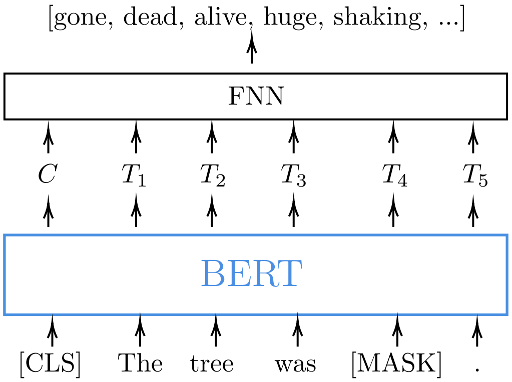


Figure 2.1: An illustrative image of BERT for masked language modelling. FNN denotes a feed-forward neural network.

i.e. predicting masked words in a text passage, and next sentence classification on a large text corpus. The training data of the BERT model consists of English Wikipedia and the Book Corpus (Zhu et al., 2015). The second training step is a fine-tuning phase during which the model can be specifically tuned to perform some kind of specific linguistic task, usually by adding a feed-forward neural network on top. With this setup, even low-resource tasks may be possible to solve thanks to the general language capabilities that have been obtained by the model in the pre-training step.

A more recent example of a model used for language representation is the RoBERTa model (Liu et al., 2019). It was developed to be more robust compared to the BERT model and is frequently used for language representation. For example, it is used in the Transformer pipeline by the text processing package spaCy.[1]

2.3 Language models

For language modelling, the goal is to generate continuations to provided input text. A *language model* (LM) expresses the probability of some next token x_n based on preceding tokens x_1, x_2, ..., x_{n-1} as follows,

$$p(x_n|x_1, x_2, ..., x_{n-1}, \theta),$$

where θ contains the parameters of the model. This format is useful for text generation, used in e.g. chatbots, question answering or story generation (Radford et al., 2018; Brown et al., 2020). Typically, models inspired by the Transformer *decoder* architecture, with causal attention and an autoregressive approach, are used as language models. It has been shown how most text-based tasks can be cast into a word completion format, meaning that there is practically no limit to what problems LMs can be applied to.

[1] `www.spacy.io`

A famous language model is the *Generative Pre-trained Transformer* (GPT) developed by Radford et al. (2018); Brown et al. (2020). It builds on the Transformer decoder and has been trained to generate the next token given preceding tokens. Differently from BERT, the GPT model cannot see the tokens after the token to be generated, in a true autoregressive fashion. More recent examples of LMs are the Pythia, Llama, Mistral and Qwen models (Biderman et al., 2023; Grattafiori et al., 2024; Jiang et al., 2023; Yang et al., 2024). These models have many properties in common with the GPT model; they mainly differ with respect to training approach and slight variations in model architecture.

Similarly to BERT, the GPT model, together with most modern language models, is pre-trained on a large text corpus to learn to compose language. For example, the GPT-3 model was trained on CommonCrawl,[2] a big dump of text from the internet, two internet-based books corpora and English-language Wikipedia (Brown et al., 2020).

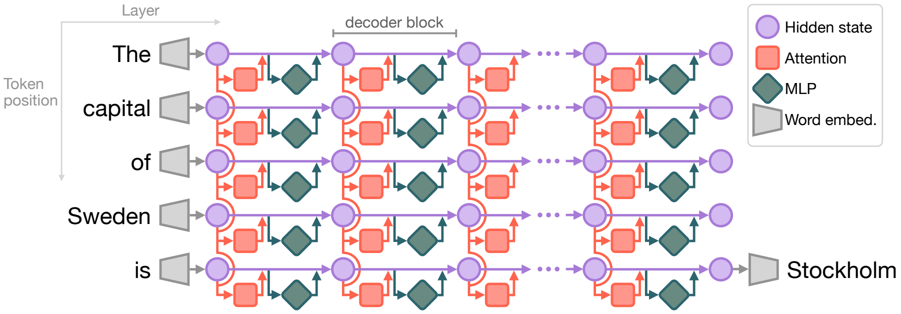


Figure 2.2: A Transformer decoder applied to a text sequence for which the computations have been rolled out. The Transformer model is applied to each token in the input. A decoder block is indicated in the figure. Word embeddings are denoted as 'Word embed.', applied either to the input tokens to encode them into numerical vectors (i.e. embeddings), or to the model output to decode numerical vectors into tokens. The figure is based on Figure 1 from Meng et al. (2022).

A closer look at the decoder part of the Transformer architecture
Some of the work included in this thesis are associated with interpretations of components of the Transformer decoder. In this section, we explain the components of the decoder in further detail. As seen in Figure 2.2, the decoder consists of *decoder blocks* stacked on top of each other. Different numbers of decoding blocks are stacked depending on the size of the decoder model, this number is typically referred to as the *number of hidden layers* of the model. In each decoder block, attention is first applied to the input, after which a *multi-layer perceptron* (MLP) is applied. Both the attention and MLP output, $\mathrm{A}(x)$ and $\mathrm{MLP}(x)$ respectively, are applied to the input x via residual

[2] `https://commoncrawl.org/overview`

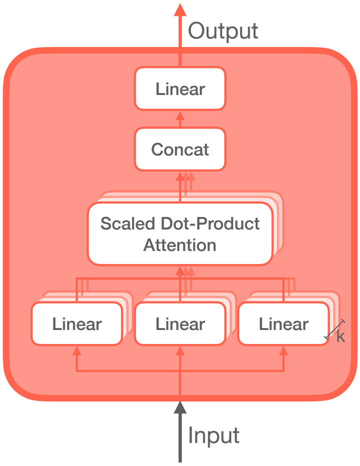


Figure 2.3: A close-up of the attention component of the Transformer with k attention heads.

connection, i.e. added in an incremental fashion to get an output y as follows.

$$y = x + \mathrm{A}(x) + \mathrm{MLP}(x)$$

The attention component of the Transformer can be decomposed into subcomponents as indicated in Figure 2.3. Most importantly, multiple *attention heads* make out the attention, allowing the model to pay attention to multiple details at the same time.

2.4 Vision-and-language models

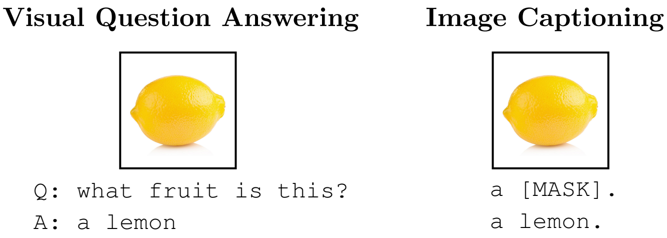


Figure 2.4: Two potential tasks for VL models. In the question answering case, the model usually generates an answer or performs a choice out of multiple options. In the image captioning case, the model can either be queried in an MLM fashion or generate a caption from scratch.

A modality that is frequently combined with text is the visual modality. Models that process both visual information and textual information are referred to as *vision-and-language* (VL) models. Examples of VL models are

VisualBERT, LXMERT, OSCAR, FLAVA and Vision Transformer (ViT) (Li et al., 2019; Tan and Bansal, 2019; Li et al., 2020; Singh et al., 2022; Dosovitskiy et al., 2021). All of these models, except for CLIP-BERT and to some extent FLAVA, have been developed to solve predominantly VL tasks, such as Visual Question Answering (VQA) or image captioning, as illustrated in Figure 2.4. Furthermore, all of these models were developed as general purpose models and can similarly to BERT be adapted to different downstream tasks.

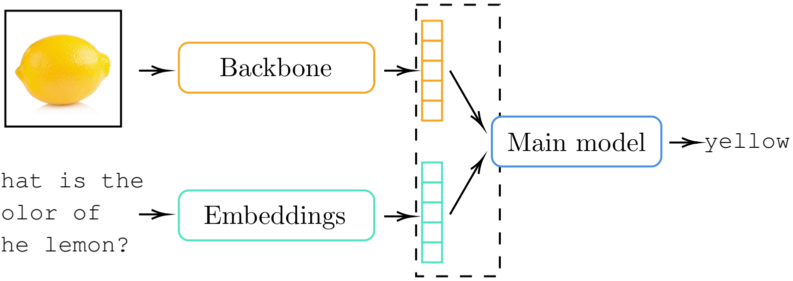


Figure 2.5: The typical setup for a VL model. Image features extracted by a backbone are given to a main model together with the text representation, usually formatted as embeddings. The dashed rectangle marks the part of the model that fuses the visual and textual information and is further described in Figure 2.6.

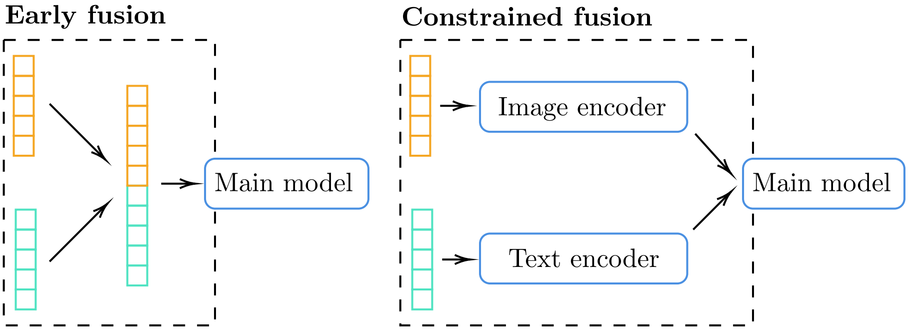


Figure 2.6: The two different fusion methods used by the VL models described in this thesis. For the early fusion, the image and text representations are simply concatenated and for the constrained fusion the representations are processed separately before the information is fused in a constrained manner through e.g. cross-attention. For the constrained fusion method the main model can also be referred to as *multimodal encoder*.

Most VL models are largely similar in their model setup, as illustrated in Figure 2.5. Typically, the models form initial representations for the visual input and textual input separately before the information from the different modalities is fused in the main model. Pre-trained word embeddings are typically used for the text input and a pre-trained visual model, generally referred to as *backbone*, is used to generate a representation for the visual input. VisualBERT, LXMERT and OSCAR use a frozen Faster R-CNN object detector (Ren et al., 2015) to extract detection features from the visual input, while CLIP-BERT utilises a frozen CLIP model (Radford et al., 2021) and FLAVA utilises a non-frozen Vision Transformer (ViT) model (Dosovitskiy et al., 2021) to generate image features. Also, all aforementioned models use Transformer encoder networks and VisualBERT, OSCAR as well as CLIP-BERT are based on a BERT model architecture.

The aforementioned VL models are also similar in their training procedure. VisualBERT, OSCAR and CLIP-BERT are initialised from pre-trained BERT-base model weights. All aforementioned VL models are then trained on image-text datasets of varying size and information content. Common for all datasets is that they either are visual question answering datasets or image captioning datasets, as illustrated in Figure 2.4. For example, VisualBERT is trained on the image captioning dataset MS COCO and the Visual Question Answering (VQA) dataset (Lin et al., 2014; Goyal et al., 2017), while LXMERT in addition to these datasets is trained on Visual Genome, GQA and VG-QA (Hudson and Manning, 2019; Zhu et al., 2016). The tasks the models are trained on differ slightly depending on model. Examples of training tasks are MLM, image-text matching and image feature prediction. Most VL models are trained on at the least MLM and image-text matching.

Chapter 3

Representations of information for natural language processing

More than linguistic knowledge is required for successful language processing. To successfully process language, commonsense world knowledge and factual knowledge is necessary. Zhang et al. (2021) suggest that the recent success of large language models on NLU benchmarks can be attributed to the capability of these models to store the required commonsense knowledge for solving the benchmarks in their *parametric memory*. In addition, Lewis et al. (2020) propose *retrieval-augmented generation* (RAG), to complement the limited knowledge of NLP models. Both the parametric memory of LMs and RAG represent different approaches to how representations of information can be used to support NLP systems (see Figure 1.1).

In this chapter, we take a closer look at representations of information for NLP systems. We consider the parametric memory – which also could be viewed as the *knowledge* of the model – (§3.1) and RAG (§3.2).

3.1 Parametric memory of language models

Most modern NLP models are *parametric models*, in the sense that they are fully described by and limited to their finite set of parameters. For some input x and potential output token y, this can be described as,

$$p(y|x, \theta, \mathcal{D}) = p(y|x, \theta),$$

where θ contains the parameters that fully describe the network and $\mathcal{D}$ is the data on which the network has been trained. Consequently, θ contains all of the information that has been learned from the training data. This is very convenient in the sense that we only need to retain the neural network parameters and not the data for subsequent text processing purposes.

Given the input "The capital of France is", most modern LMs are capable of generating the correct completion "Paris". LMs are considered to have a *parametric memory* capable of storing factual information related to e.g. capitals (Petroni et al., 2019). It is not fully known what is stored in this memory and how it is impacted by training data. Recent research has found LMs more likely to have memorised *popular facts*, expected to have occured more frequently in the training data (Mallen et al., 2023). Zhang et al. (2021) also hypothesise that LMs require large amounts of training data to learn the necessary knowledge for successfully processing natural language.

We may refer to the information represented by the parametric memory of an ANN as *parametric knowledge*. However, we consider this knowledge to be different to that of a human, since it cannot be expected to have the same properties. For example, it is not known whether the parametric memory can robustly represent information without contradiction, something that generally can be expected of human knowledge (Brachman and Levesque, 2004). For example, Elazar et al. (2021) probe the factual consistency of LMs, finding that they are sensitive to insignificant changes in queries for factual information, indicating that the parametric memory of LMs can be contradictory, or at least difficult to access in a consistent manner.

3.2 Retrieval-augmented generation

External representations of information, as opposed to the internal parametric memory, can also be used to support NLP systems. Lewis et al. (2020) argue that the ability of LMs to access and precisely manipulate their parametric knowledge is limited, making the models underperform on knowledge-intensive tasks. As an alternative, Lewis et al. (2020) propose to use retrieval-augmented generation (RAG) – models which combine pre-trained parametric and non-parametric memory for language generation. This can be described as,

$$p(y|x, \theta) = p(y|x, R(x), \theta),$$

where x denotes the input, y some output token, and θ the model parameters. $R(x)$ denotes additional external information that has been *retrieved* based on the input. θ thus represents the parametric memory and $R(x)$ the non-parametric memory.

By now, many different RAG systems have been proposed and evaluated (Gao et al., 2024). While the design of the systems may vary, they all share certain features, see Figure 3.1. All RAG systems are based on a retrieval corpus, which may cover any form of representation(s) of information (e.g. an unstructured text dump, a knowledge graph, or both), combined with an automated information retrieval component that, given a query, is capable of fetching relevant information from the retrieval corpus. Furthermore, all systems include some form of information fusion step, at which the retrieved information is incorporated for the LM output, together with the query. Initial RAG research typically involved quite advanced fusion methods, such as *fusion-in-decoder* (Izacard and Grave, 2021). Nowadays, most methods simply provide the retrieved information in the input, prepended to the query.

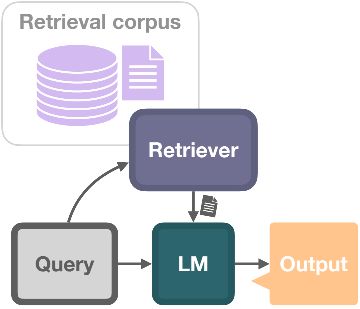


Figure 3.1: An illustration of a typical RAG design. A query is passed to both the retriever and LM (sometimes referred to as 'reader'). The retriever then fetches relevant entries from the retrieval corpus and provide these to the LM. Finally, the LM generates an output based on the query and retrieved information.

Examples of LMs used for RAG are Atlas (Izacard et al., 2023) and standard LMs such as the Llama and Qwen models (Grattafiori et al., 2024; Yang et al., 2024). Atlas has been specifically developed and tuned for RAG applications, the same does not hold for standard LMs. In spite of this, standard LMs, for which the retrieved information is simply prepended to the input in a zero-shot fashion, have been found to work well in RAG systems (Ram et al., 2023). This is attributed to the strong generalisation abilities of modern LMs.

3.2.1 Utilisation of external information

Much of the work included in this thesis is focused on the ability of LMs to utilise external information, also referred to as *context utilisation*. Context utilisation is a key component of LMs used for RAG, as the benefits of retrieving external information are realised only if the generative model makes adequate use of the retrieved information. In this section, we consider challenges related to context utilisation and methods for improving context utilisation.

Many weaknesses of LMs used for RAG are associated with context utilisation. For example, LMs can easily be distracted by irrelevant contexts (Shi et al., 2023a) or ignore relevant contexts due to memory-context conflicts (Xu et al., 2024). The robustness of LMs to irrelevant contexts is important as information retrieval systems used for RAG are not guaranteed to always retrieve relevant information. Moreover, as information may be updated to conflict with the training data of the LM, the model should prioritise the most recently updated information.

As a consequence, many context utilisation manipulation techniques (CMTs) have recently been proposed to improve LM context utilisation. Existing CMTs can be categorised into one of four main groups based on *intervention level*, i.e. what aspect of the model they manipulate. 1) *fine-tuning* CMTs update model parameters to modify context utilisation. For example, fine-tuning on distracting contexts was found to yield improved robustness to distracting contexts (Li et al., 2023; Shen et al., 2024; Yoran et al., 2024). Moreover, Fang et al. (2024) specifically focus on different types of retrieval noise likely to be encountered in real-world environments and develop a fine-tuning approach to handle these. 2) *prompting techniques* modify the input to the LM to improve context utilisation, representing minimally modified settings. 3) *mechanistic interventions* on the LM modify certain model components at inference time to alter context utilisation. Examples involve attention modification (Ortu et al., 2024; Jin et al., 2024) and SpARe interventions (Zhao et al., 2025). Lastly, 4) *decoding methods* involve a modified decoding approach, applied to the output logits, to manipulate context utilisation. Examples include context-aware contrastive decoding (Yuan et al., 2024; Kim et al., 2024; Shi et al., 2024; Wang et al., 2024a; Zhao et al., 2024) and lookback lens decoding (Chuang et al., 2024).

Apart from intervention level, many of the CMTs have different *objectives*, focused on improving one or multiple aspects of context utilisation. CMTs may focus on improving robustness to irrelevant contexts, faithfulness to conflicting contexts, or faithfulness to contexts in general.

Chapter 4

Knowledge-centered evaluations of systems for natural language processing

As explained in the introduction, much of the work included in this thesis is focused on developing sound evaluation methods for knowledge-intensive tasks. In this chapter, we further expand on the necessity of proper evaluation methods (§4.1) and describe current approaches for measuring the parametric knowledge of LMs (§4.2) as well as their ability to leverage external information (§4.3).

4.1 Evaluating systems for natural language processing

All NLP models based on ANNs need to be carefully evaluated to assert that they work as intended, since they risk learning to make correct inferences with an incorrect method. Very few modelling constraints are necessary or imposed on the models to guide their inference process, as this is also a strength of neural networks. However, this means that the performance of ANNs only can be measured via empirical approaches rooted in evaluation data that has been designed to elicit and test the trait under consideration. This in turn means that the models may learn to make use of spurious correlations or artifacts in the training data to solve the evaluation data and not the task (Feder et al., 2022). NLP models have been noted to base predictions that should be factual on the style of their prompts, meaning that they might switch their prediction when they are queried for the same fact but in a different way (Elazar et al., 2021; Cao et al., 2021; Jiang et al., 2020). Additionally, NLP models solve the grounded commonsense inference task SWAG by picking up on dataset-specific distributional biases instead of learning to perform commonsense inference (Zellers et al., 2018, 2019). It has also been observed that BERT picks up on

Astrid Lindgren was a Swedish writer and children's book author best known for her character Pippi Longstocking. She was born on November 14, 1907 in Vimmerby, Sweden. Lindgren grew up in a family with strong storytelling traditions, and she began writing stories and poems from a young age.

After completing her education, Lindgren worked as a secretary and later as a journalist. In the 1930s, she married Sture Lindgren and had two children. In the 1940s, she began writing children's books, starting with the character Pippi Longstocking, which became her most well-known and beloved creation.

Lindgren's books have been translated into more than 70 languages and have sold millions of copies worldwide. In addition to Pippi Longstocking, Lindgren is also the author of numerous other children's books, including the Ronia the Robber's Daughter series, The Children of Noisy Village, and Karlsson-on-the-Roof.

Lindgren was also a strong advocate for children's rights and was involved in various humanitarian and political causes throughout her life. She received numerous awards for her work, including the Hans Christian Andersen Award for Children's Literature in 1958 and the Right Livelihood Award in 1979. Lindgren passed away on January 28, 2002 at the age of 94.

Figure 4.1: Hallucination in ChatGPT. The model was prompted with the query "Give me the biography of Astrid Lindgren.". Passages marked in yellow are questionable or clearly incorrect.

shallow heuristics such as lexical overlap to solve a general natural language inference task (McCoy et al., 2019).

Generative NLP models also suffer from *hallucinations*, which may be particularly difficult to detect (Maynez et al., 2020; Shuster et al., 2021). As illustrated in Figure 4.1, even sophisticated models such as ChatGPT may generate plausible but incorrect facts. The text generated by the model states that Lindgren received the Right Livelihood Award in 1979, while she actually received it in 1994. If these models are used for low-risk scenarios such as storytelling, there is no immediate danger. It is, however, a problem for situations in which correct information is important and expected.

4.2 Evaluating for parametric knowledge

There is much interest in extracting and measuring the different types of knowledge that supposedly resides in NLP models (i.e. the *parametric memory*). For BERT-like models, sentence completion tasks, also known as cloze statements, are typically used to evaluate knowledge since the models are tuned to this format from their MLM pre-training. Petroni et al. (2019) use this format to test for factual and commonsense knowledge with their LAMA (LAnguage Model Analysis) probe based on Wikipedia and commonsense knowledge. They find that much knowledge is stored in language models and further hypothesise that these models have a potential use as knowledge bases. For GPT-like models, a similar approach can be applied, asking the model to generate the continuation that completes a fact. Weir et al. (2020) also test for commonsense knowledge, such as "A dog has fur.", and find that these are present in different

BERT based models. West et al. (2022) experiment with extracting latent commonsense knowledge from a GPT-2 model to create a knowledge graph, and use different prompting techniques for this.

Orthogonal to this, Elazar et al. (2021) consider the *factual consistency* of LMs. Situations related to factual knowledge require not only high accuracy but also consistency, i.e. robustness to lexical variations in semantically equivalent queries. Recent LM developments have mainly improved on accuracy, while the question of consistency has seen less attention. Elazar et al. (2021) find that even SoTA models may produce different outputs depending on lexical variations in semantically equivalent queries.

Apart from black box model evaluations focused only on the model output, there is also work that tries to open the black box to perform a deeper inspection of the parametric knowledge of LMs. Meng et al. (2022); Geva et al. (2023) *locate* stored information in LMs and inspect how it *flows* through the model to finally be expressed in its output. They focus on the inference process of LMs for fact completion for simple ⟨subject, relation, object⟩ fact tuples, such as subject `Tokyo`, relation `capital_of` and object `Japan` (e.g. for the query "What is the capital of Japan?"). To enable these investigations, it is first asserted that the model under consideration has the necessary knowledge, such that it can be located in the model parameters. This is done by querying the model for different facts (e.g. "What is the capital of Japan?"), and checking if the model gets the facts correct. It is assumed that the model has the factual information stored in its parameters if it correctly answers the corresponding query.

4.3 Evaluating for the utilisation of external information

LMs used for knowledge-intensive tasks or RAG need be good at leveraging external information, i.e. have good context utilisation. A large body of recent research has focused on evaluating the context utilisation of LMs using *context-intensive datasets*, i.e. datasets representing tasks that are difficult to solve without good context utilisation. The samples in these datasets usually contain both a query and the corresponding context, dropping the information retrieval step in RAG to be able to control for context characteristics and their impact on context utilisation.

We consider two main categories of context-intensive datasets: 1) datasets representing *knowledge-intensive tasks*, i.e. tasks for which access to external context is crucial, and 2) datasets designed to *diagnose* model adaptability to external information. Examples of knowledge-intensive datasets representative of the former category are Natural Questions (NQ), the KILT datasets and PubMedQA (Kwiatkowski et al., 2019; Petroni et al., 2021; Jin et al., 2019).

Examples of diagnostic datasets representative of the latter category are CounterFact and ConflictQA (Meng et al., 2022; Xie et al., 2024). These datasets contain synthesised queries based on fact triplets from LAMA (Petroni et al., 2019) (e.g. ⟨`Thomas Ong`, `citizen_of`, `Singapore`⟩) for which contexts have been synthesised to induce *knowledge conflicts* by promoting answers in conflict with the parametric memory of the studied LM (e.g. `Pakistan` as opposed to `Singapore`). Diagnostic datasets have found widespread use for work on mechanistic interpretability and the evaluation of context utilisation (Meng et al., 2022; Geva et al., 2023; Ortu et al., 2024).

Work in information retrieval and RAG has identified several qualities in retrieved or synthesised contexts that impact context utilisation by humans and/or LMs. Retrievers typically provide overly long or corrupted text, which are *difficult to understand*, and impact LM output (Gao et al., 2024; Vladika and Matthes, 2023). Similarly, typos (Cho et al., 2024) and high perplexity (Gonen et al., 2023) have been identified as potential disruptors for RAG systems. Furthermore, *implicit* contexts, lacking an explicit connection to the query, have been identified as a prevalent failure cause in RAG (Li et al., 2024). For automated retrieval situations, the rate of implicit contexts can be high due to chunking of text (Wang et al., 2024b). Instead, LMs have been shown to prefer context with high *query-context similarities* (Wan et al., 2024).

Most studies on RAG have focused on open-domain question answering (Kasai et al., 2023; Wu et al., 2024). Yoran et al. (2024); Shi et al. (2023b) found that LMs are fragile to *irrelevant information* in the context, harming performance. Furthermore, in the case of *knowledge conflicts*, when context conflicts with parametric knowledge, LMs have been shown to ignore the conflicting context (Longpre et al., 2021), while other studies show that models prefer contextual information, as long as it is coherent and convincing (Xie et al., 2023). Sun et al. (2025) also connect knowledge conflicts to prediction uncertainty in fact-checking settings. Recently, Xu et al. (2024) have proposed more granular categories for knowledge conflicts, using *context-memory conflict* to denote the aforementioned phenomenon, and *inter-context conflict* to refer to different contexts contradicting each other.

Unreliable contexts have been studied by Chrysidis et al. (2024) in a fact-checking setup, for which misinformation is prevalent. This type of information is typically overlooked in more generic RAG QA setups, potentially because the retrieval corpora usually are based on Wikipedia or pre-curated datasets. *References to external sources* may convince a human reader of the credibility of some context, yet LMs seem to be unaffected by references (Wan et al., 2024). However, expressed *certainty/uncertainty* in text and its impact on LM context usage has recently been studied by Du et al. (2024a), where assertive contexts are found to be more convincing.

Chapter 5

Summary of included papers

The papers in this thesis have mainly focused on the intersection between LMs, knowledge and representations of information. LMs used for knowledge-intensive tasks need not only be accurate, but also factually consistent, updatable and, ultimately, reliable (RQ1). To elicit and measure these aspects of interest in knowledge-intensive situations, we also need appropriate evaluation data and evaluation methods (RQ2). RQ1 is addressed in papers 1, 2, 4 and 5. RQ2 is addressed in papers 1, 3, 4 and 5.

5.1 Paper 1: Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?

Paper 1 studies and evaluates the acquisition of knowledge related to visual concepts (such as the colour of different well-known items). We investigate the use of visual data to complement the knowledge of large language models and propose a method for evaluating visual knowledge transfer to text and introduce a novel text-only task, Memory Colors, querying for knowledge of memory colours, i.e. typical colours of well-known objects (Pérez-Carpinell et al., 1998). The task is in English and contains 109 object types paired with their memory colour according to the knowledge of 11 human annotators. An example of a query from Memory Colors is "What is the color of a lemon? [MASK]", where [MASK] should be filled in with the correct answer (yellow), see Figure 5.1.

Similarly to the case for humans, we assume that a model with sufficient knowledge of visual concepts should be able to answer text-only queries about their colours without necessarily being provided with images of the concepts. To support this point, we complement Memory Colors with a human baseline from 11 human annotators that did not have access to images while answering

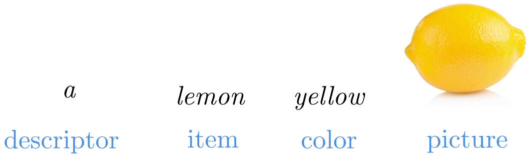


Figure 5.1: One entry in the Memory Colors dataset. When evaluating a model on Memory Colors, the descriptor and item are slotted into a pre-defined template, e.g. “What is the color of [descriptor] [item]?”.

the queries.

We also introduce a novel VL model architecture, CLIP-BERT that utilises CLIP as backbone and BERT-base as main model. We train it on 4.4M captions and 2.7M images. After training, it can be used to make inferences in an implicit or explicit mode. In the implicit mode, the model is queried with only text and in the explicit mode it is also provided with a visual representation of the text generated by CLIP. Since CLIP has been trained to map visual and textual representations to the same space, it has a potential use for “imagining” a visual representation corresponding to text when no image is available. We also measure an upper bound for the CLIP-BERT performance by evaluating it when it is provided with images corresponding to the text.

To separate and investigate the knowledge contributions from text versus images, we experiment with removing information about visual concepts from the text part of the training data by using different filtering methods. For example, we might remove a training example from the data if it contains an object and its corresponding colour from Memory Colors. In this way we can clearly separate knowledge contributions from images and text respectively.

Finally, we evaluate CLIP-BERT on Memory Colors in the different modes with the different filterings of the training data. We also evaluate its text-only counterpart, BERT-base, trained on the same different filterings of the training data. We find that CLIP-BERT outperforms BERT in every filtering setting, and with a larger margin if visual information is filtered out from the text data used for training. We also find that a CLIP-BERT model in explicit mode has a larger performance margin when visual information has been removed from the training corpus. This indicates that our method can successfully be used to measure visual knowledge transfer capabilities in models and that our novel model architecture shows promising results for leveraging multimodal knowledge in a unimodal setting.

Taken together, Paper 1 provides insights related to how uni- and multimodal LMs acquire and utilise visual knowledge. To enable these insights, the paper introduces an evaluation dataset that measures visual knowledge, Memory Colors. A noteworthy finding of the paper is related to prompt sensitivity; model performance is greatly influenced by the phrasing of the query, more so than for humans. This raises the question of how to handle measurements of *knowledge* of a LM, when it is subject to change from insignificant changes in the query, further studied in Paper 2.

Contributions T. Norlund mainly contributed to the design of the study, implemented the CLIP-BERT model and code for evaluating it. He also made major contributions to the writing of the paper.

L. Hagström mainly contributed to the design of the study and developed the Memory Colors dataset. She also made major contributions to the writing of the paper.

R. Johansson provided supervision on the work and writing for the paper.

5.2 Paper 2: The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models

Paper 2 moves away from the visual and multimodal domains, sharpening the focus on evaluating LMs for factual knowledge. Given that LMs are increasingly applied to knowledge-intensive tasks, it is important that they are consistent and robust to insignificant changes in their input. However, Elazar et al. (2021) found that modern LMs have poor factual consistency. For example, the same LM may predict "Anne Redpath's life ended in *London*" and "Anne Redpath passed away in *Edinburgh*". In this paper, we further investigate the factual consistency of LMs and explore methods for improving it, see Figure 5.2.

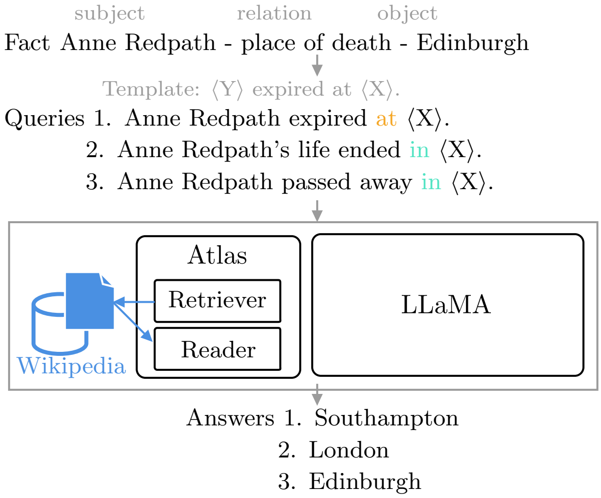

subject
relation
object
Fact Anne Redpath - place of death - Edinburgh
Template: ⟨Y⟩ expired at ⟨X⟩.
Queries 1. Anne Redpath expired at ⟨X⟩.
2. Anne Redpath's life ended in ⟨X⟩.
3. Anne Redpath passed away in ⟨X⟩.
Atlas
Retriever
Reader
Wikipedia
LLaMA
Answers 1. Southampton
2. London
3. Edinburgh

Figure 5.2: Overview of how consistency is computed in ParaRel for Atlas and LLaMA.

We evaluate the effectiveness of two mitigation strategies for improved factual consistency; up-scaling to larger model sizes and augmenting the LM with a retrieval corpus. Increasing the size of LMs has previously been shown to work well as an multi-purpose tool for improving most aspects of model performance. More recently, another approach has been proposed for improving

LM performance: change the model design itself to be guided by inductive biases that promote various desirable properties. Examples of such models are text retrieval-augmented models that condition predictions on retrieved text passages for improved adaptability, interpretability and efficiency, also referred to as *retrieval-augmented generation* (RAG) (Lewis et al., 2020).

To evaluate effects of upscaling on factual consistency, we study the performance of Llama models of sizes 7B, 13B, 33B and 65B parameters (Touvron et al., 2023). To evaluate effects of retrieval augmentation, we study the performance of Atlas-base and Atlas-large, two retrieval-augmented models corresponding to 330M and 880M parameters, respectively, augmented with text passages retrieved from Wikipedia (Izacard et al., 2023).

To measure the factual consistency of LMs we use an improved version of ParaRel, denoted ParaRel*. ParaRel was originally developed by Elazar et al. (2021) and is improved by us via the removal of duplicated samples and addition of four query-level metrics to estimate query related inconsistency sources. ParaRel is based on LAMA (Petroni et al., 2019), an evaluation task based on Wikidata that measures factual knowledge stored in LMs through prompting for the missing object given a subject-relation tuple. ParaRel adds a layer of semantically equivalent cloze-style prompts to LAMA, which in turn allows us to measure the consistency of LMs with respect to the knowledge triples represented by LAMA (see Figure 5.2). The idea is that a model is consistent if it is invariant to query paraphrases.

Evaluations of the Llama and Atlas models on ParaRel* reveal that both upscaling and retrieval-augmentation improve factual consistency. However, retrieval-augmentation is found to be more efficient compared to upscaling; the Atlas-base model performs on par with the Llama 65B model despite being 90 times smaller.

To better understand *why* and *when* LMs can be expected to be factually (in)consistent, we investigate potential causes of inconsistency. We find that different aspects of form impact consistency. High lexical similarity between subject and object (the correct answer) generally leads to improved consistency. Samples for which the correct prediction would produce an unidiomatic sentence (e.g. "Solar Mass is named after *Sun*" as opposed to "Solar Mass is named after *the Sun*") are also found to correspond to lower consistency. Altogether, these results show how LMs may prioritise correct form over consistency, which may not be entirely surprising as LMs have previously been found to learn form and syntax faster than semantics and general natural language understanding (Zhang et al., 2021). This also raises questions related to the effects of using synthesised datasets, e.g. by slotting fact tuples into prompt templates with the risk of producing less fluent sentences, to probe for model capabilities. How can we be certain that our findings reflect the phenomenon of interest, as opposed to consequences of data artifacts?

We further investigate the consistency of Atlas, focused on effects of the retriever. We find that retriever consistency is weakly correlated with Atlas consistency and that consistent and relevant retrieval causes more consistent predictions. However, not even for perfectly consistent retrieved passages does Atlas achieve perfect consistency, indicating that more inconsistency sources are involved and persistent in spite of consistent conditioning.

Altogether, Paper 2 evaluates the effectiveness of different types of LMs in a fact-intensive setting, for which factual consistency is important. We find that RAG excels in these settings, corroborating previous work which has found RAG to work well for fact-intensive QA (Izacard et al., 2023; Lewis et al., 2020) and reduced hallucination (Shuster et al., 2021; Thoppilan et al., 2022). Paper 2 thus provides further guidance on how to obtain reliable NLP systems for knowledge-intensive situations.

Contributions L. Hagström implemented and evaluated the Atlas models. She also contributed to the development of ParaRel* and 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 development of ParaRel* and 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.

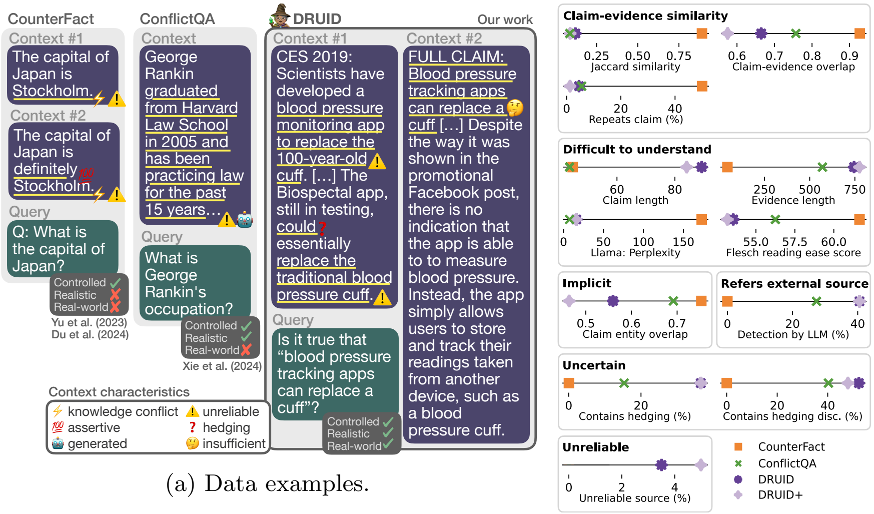
M. Johansson provided supervision on the work and writing of the paper.

R. Johansson provided supervision on the work and writing of the paper. He also helped write the paper.

5.3 Paper 3: A Reality Check on Context Utilisation for Retrieval-Augmented Generation

As observed in Paper 2, RAG can be used to alleviate different problems arising from the imperfect parametric knowledge of language models (LMs), which may encode unstable, limited or potentially outdated information (Gao et al., 2023; Vu et al., 2024). However, as also found in Paper 2, the benefits of RAG are only realised if 1) the retrieval module retrieves helpful information and 2) the generative model successfully leverages the retrieved information. Paper 3 takes a closer look at the latter aspect, introducing DRUID (Dataset of Retrieved Unreliable, Insufficient and Difficult-to-understand contexts) to facilitate investigations into context utilisation in real-world scenarios.

Previous work on context utilisation has mainly studied RAG in a disjoint manner, where studies of the quality and relevance of the retrieved information are detached from studies of LM context usage (Shi et al., 2023b; Xie et al., 2023; Tan et al., 2024; Du et al., 2024a) (see Figure 5.3(a)). Hence, little is understood about 1) the characteristics of retrieved contexts and 2) their impact on LM context usage (see Figure 5.3(a)). Previous studies of context utilisation have mainly been based on synthesised datasets like CounterFact (Ortu et al., 2024) and ConflictQA (Xie et al., 2024), most likely since these are easy to obtain and control. However, the scenarios described by these datasets are not representative of *real-world* RAG scenarios, as the context types do not reflect the diversity and complexity of the ones returned by an actual retriever



CounterFact

Context #1: The capital of Japan is Stockholm. ⚡⚠

Context #2: The capital of Japan is definitely 💯 Stockholm. ⚡⚠

Query: Q: What is the capital of Japan?

Controlled ✓ Realistic ✗ Real-world ✗

Yu et al. (2023) Du et al. (2024)

ConflictQA

Context: George Rankin graduated from Harvard Law School in 2005 and has been practicing law for the past 15 years… ⚠🤖

Query: What is George Rankin's occupation?

Controlled ✓ Realistic ✓ Real-world ✗

Xie et al. (2024)

DRUID Our work

Context #1: CES 2019: Scientists have developed a blood pressure monitoring app to replace the 100-year-old ⚠ cuff. […] The Biospectal app, still in testing, could ❓ essentially replace the traditional blood pressure cuff. ⚠

Context #2: FULL CLAIM: Blood pressure tracking apps can replace a 🤔 cuff […] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.

Query: Is it true that "blood pressure tracking apps can replace a cuff"?

Controlled ✓ Realistic ✓ Real-world ✓

Context characteristics

⚡ knowledge conflict ⚠ unreliable
💯 assertive ❓ hedging
🤖 generated 🤔 insufficient

(a) Data examples.

(b) Context characteristics.

Figure 5.3: Comparisons between ConflictQA, CounterFact and DRUID.

present in RAG (Longpre et al., 2021; Ravaut et al., 2024; Ortu et al., 2024). DRUID aims to address this.

To create DRUID, we focus on the prototypical information-seeking task of fact verification, where retrieving and utilising real-world information is vital. For the task, an agent is provided with a statement about the world – a *claim* – and needs to decide whether it is true or false using context retrieved from an external source – *evidence* (Guo et al., 2022). We take real fact-checked claims as 'queries' and evidence retrieved from the web by an automated information retrieval pipeline as 'context' to evaluate RAG in this real-world setting, which naturally facilitates our goal of studying real-world context properties in RAG (Samarinas et al., 2021; Atanasova et al., 2022; Chrysidis et al., 2024; Glockner et al., 2024). To assess the relevance and stance of the retrieved evidence, necessary for studies of context utilisation, we crowd-source evidence-level annotations. A DRUID sample consists of a ⟨claim, evidence, labels⟩ triple, and we collect a total of 5,490 double-annotated samples.

To understand the gap between the context provided in synthesised diagnostic datasets for context utilisation and real RAG scenarios, we compare the characteristics within DRUID to the synthesised CounterFact and ConflictQA datasets. We consider context characteristics previously observed to impact context utilisation by humans and/or LMs. These are related to the stance of the context, query-context similarity, whether the context is difficult to understand, whether the context only implicitly refers to the query, whether the context refers to an external source, and whether the context is uncertain or unreliable. We find that the real-world samples in DRUID contain many contexts (50%) that are *insufficient*, i.e. they are relevant to the query but do not contain sufficient information to answer the query, compared to Counter-

Fact and ConflictQA for which only sufficient contexts are found. For the other context characteristics considered, we also find a great discrepancy between CounterFact, ConflictQA and DRUID, further proving the need of datasets aligned with real-world RAG scenarios like DRUID (see Figure 5.3(b)).

We also evaluate LM context utilisation on DRUID and compare this to insights based on synthesised datasets. To this end, we measure the context utilisation of Pythia 6.9B and Llama 3.1 8B on the CounterFact, ConflictQA and DRUID datasets. To measure context utilisation we introduce the ACU metric that compares the softmaxed-normalised model logits for inputs without context and with context. ACU values lie in the range $[-1, 1]$, for which a value of 1 indicates perfect context usage and a value of -1 indicates *context repulsion*, for which the LM output fully contradicts the context. We find that synthetic datasets suggest an over-preference of supporting evidence, and that context utilisation differs between the LMs studied, for which Llama generally is better at utilising contexts compared to Pythia.

Finally, we evaluate the influence of different context characteristics on model context usage. For this, we calculate Spearman correlations between each context property and our context usage metric, ACU, stratified by the evidence stance for each dataset. We find that contexts from fact-check sources correspond to greater ACU scores, that references to external sources show low correlations with ACU and that correlations with query-context similarity are low for DRUID while they are high for ConflictQA. Our results indicate that real-world queries and contexts come with a greater complexity for which context usage cannot be predicted solely based on e.g. query-context similarity.

To summarise, Paper 3 grounds studies of context utilisation to real-world RAG scenarios. It also provides deeper insights related to how LMs use context of different characteristics. As a result, our understanding of how LMs interact with external knowledge representations, here represented by passages retrieved from the web, is improved.

Contributions L. Hagström was the project leader, designed the automated retrieval pipeline for context collection, collected the manual annotations for DRUID, evaluated the LMs under consideration, performed the context characteristics analysis and contributed to the writing of the paper.

S.V. Marjanović helped design the annotation guidelines and contributed to the design of the context utilisation metric. She also consulted on the general design of the method and made major contributions to the writing of the paper.

H. Yu helped design the annotation guidelines and contributed to the automated retrieval of contexts. She also consulted on the general design of the method and contributed to the writing of the paper.

A. Arora helped design the annotation guidelines. He also consulted on the general design of the method and made major contributions to the writing of the paper.

C. Lioma provided supervision on the work and writing of the paper.

M. Maistro provided supervision on the work and writing of the paper. She also provided comprehensive advice for the annotation collection and automated retrieval of contexts.

P. Atanasova helped design the annotation guidelines and made major contributions to the implementation of the annotation platform and the collection of the annotations. She also consulted on the general design of the method and made major contributions to the writing of the paper.

I. Augenstein helped design the annotation guidelines and provided supervision on the work and writing of the paper. She also made major contributions to the writing of the paper.

5.4 Paper 4: Fact Recall, Heuristics or Pure Guesswork? Precise Interpretations of Language Models for Fact Completion

Papers 2 and 3 have mainly studied how LMs interact with *external* knowledge representations used in retrieval-augmented generation. In Paper 4 we instead study the *internal* knowledge representations of LMs, i.e. the knowledge LMs have acquired from pre-training on vast corpora.

LMs have been found to store significant amounts of factual information (Petroni et al., 2019). While there are many research results documenting the fact proficiency of LMs (Kandpal et al., 2023; Mallen et al., 2023), our understanding of how these models perform fact completion is still under development. Mechanistic interpretability is a growing area of research aiming to explain model behaviour (Elhage et al., 2021; Geiger et al., 2021), and has already yielded insights into where LMs store and process factual information for accurate predictions (Meng et al., 2022; Geva et al., 2023; Haviv et al., 2023).

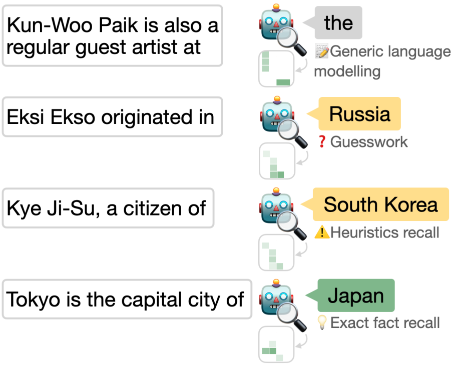


Figure 5.4: Prediction scenarios and corresponding prompt completion examples. Each scenario yields distinct interpretability results.

With Paper 4, we expand on the scenarios studied in mechanistic interpretability, extending the analysis to cover four scenarios (generic language modelling, guesswork, heuristics recall, and exact fact recall) as opposed to one (accurate prediction), see Figure 5.4. We hypothesise that the ‘accurate prediction' scenario studied in previous work in reality is a blend of multiple fine-grained scenarios, as it is well known that LMs can make correct predictions based on many different signals in the prompt, not all corresponding to *exact fact recall* (Zellers et al., 2019; Niven and Kao, 2019; McCoy et al., 2019; Poerner et al., 2020; Cao et al., 2021; Ladhak et al., 2023). The four disentangled prediction scenarios identified in our work are defined as follows: 1) *Generic language modelling*, when the model does not respond with facts, such as when generating a story. 2) *Guesswork*, when the model responds with a fact but is uncertain. 3) *Heuristics recall*, when the model uses shallow heuristics, e.g. that people with Korean-sounding names are more likely to live in Korea. 4) *Exact fact recall*, when the model has indeed memorised the correct answer and recalls it for the prediction.

We propose the PrISM method for creating a diagnostic dataset with distinct test cases. The method is based on three necessary and comprehensive diagnostic criteria for which we define measurements: (1) Does the prediction represent fact completion rather than generic language modelling? (2) Is the prediction confident and robust to insignificant signals in the prompt? (3) Is the prediction based on the exact factual information expressed in the query or on heuristics triggered by surface-level cues? These criteria provide a more fine-grained testing setup compared to using a single accuracy-focused criterion. Using the criteria, we build PrISM datasets with ⟨*query*, *prediction*⟩ samples representative of each of the four prediction scenarios, following the process described in Figure 5.5.

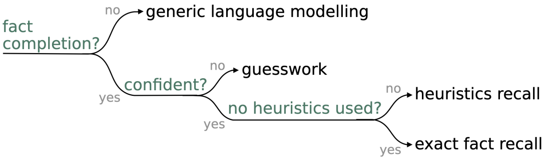


Figure 5.5: Diagnostic criteria (in green) for defining the four prediction scenarios (in black).

To test whether the proposed four prediction scenarios yield different interpretability results, we apply two mechanistic interpretability approaches – causal tracing (CT) (Meng et al., 2022) and information flow analysis (Geva et al., 2023) – to LMs evaluated on PrISM. We find that different prediction scenarios yield distinct interpretability results if studied in isolation, while model interpretations over the 'accurate prediction' scenario yield averaged and imprecise results in comparison. Our results corroborate and clarify previous insights related to how LMs process factual queries. We also provide new insights related to how LMs process factual information for heuristics recall and guesswork scenarios.

To summarise, Paper 4 facilitates precise interpretations of LMs by expanding on and delineating fact completion scenarios for which we can interpret LMs. Consequently, it helps to improve our understanding of how LMs leverage the internal knowledge representations found in their model parameters.

Contributions D. Saynova contributed to the identification method of prediction scenarios, the creation of the PrISM datasets 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, the creation of the PrISM datasets 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.

5.5 Paper 5: CUB: Benchmarking Context Utilisation Techniques for Language Models

Paper 5 largely builds on Paper 3 by taking a closer look at LM context utilisation. Specifically, it develops CUB (Context Utilisation Benchmark) to allow for a comprehensive evaluation and comparison of context usage manipulation techniques (CMTs), see Figure 5.6.

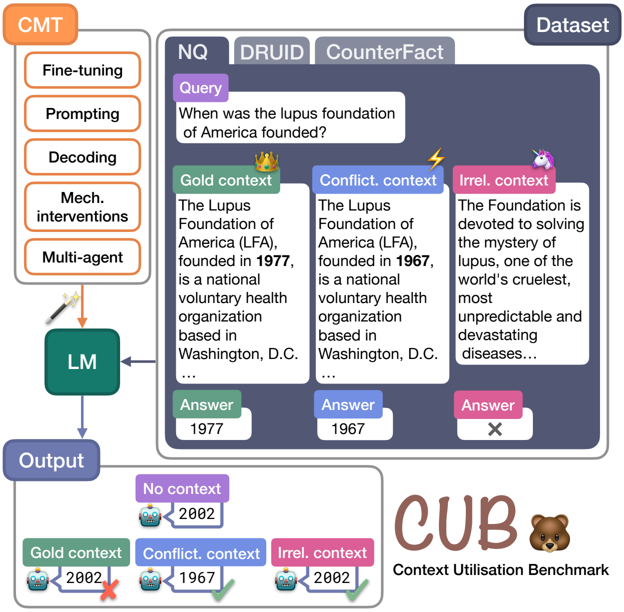

CMT
Fine-tuning
Prompting
Decoding
Mech. interventions
Multi-agent
LM
Output
Dataset
NQ
DRUID
CounterFact
Query
When was the lupus foundation of America founded?
Gold context
The Lupus Foundation of America (LFA), founded in 1977, is a national voluntary health organization based in Washington, D.C. ...
Conflict. context
The Lupus Foundation of America (LFA), founded in 1967, is a national voluntary health organization based in Washington, D.C. ...
Irrel. context
The Foundation is devoted to solving the mystery of lupus, one of the world's cruelest, most unpredictable and devastating diseases...
Answer
1977
Answer
1967
Answer
✕
No context
2002
Gold context
2002
Conflict. context
1967
Irrel. context
2002
CUB
Context Utilisation Benchmark

Figure 5.6: The Context Utilisation Benchmark. We evaluate a range of LMs under different CMTs on samples from NQ, DRUID and CounterFact for gold, conflicting and irrelevant contexts.

Context utilisation is a key component of LMs used for RAG, as the benefits of retrieving external information are only realised if the generative model makes adequate use of the retrieved information. While recent research has identified many benefits of augmenting LMs with retrieved information (Shuster et al., 2021; Hagström et al., 2023), it has also identified weaknesses of LMs used for RAG, of which many are associated with context utilisation. For example, LMs can easily be distracted by irrelevant contexts (Shi et al., 2023a) or ignore relevant contexts due to memory-context conflicts (Xu et al., 2024). As a consequence, many different methods for increasing or suppressing LM context utilisation have been proposed. The methods encompass a broad range of approaches (Shi et al., 2024; Kim et al., 2024; Li et al., 2023; Liu et al., 2023; Feng et al., 2024; Du et al., 2024b; Ortu et al., 2024; Jin et al., 2024). While each method yields promising results in isolation, their evaluation is often limited to narrow or idealised settings, leaving open the question of which approaches are applicable in real-world RAG scenarios.

To address this knowledge gap, CUB systematically tests the sensitivity of CMTs to underlying model and naturally occurring context types (gold, conflicting and irrelevant) on tasks representative of synthesised and realistic RAG scenarios. To evaluate the model sensitivity of CMTs, CUB perform evaluations on up to nine different LMs. To evaluate how CMTs respond to different types of contextual information, CUB evaluates each CMT on CounterFact (Meng et al., 2022), NQ (Longpre et al., 2021) and DRUID. The inclusion of these datasets is based on three key criteria: (i) diversity in task difficulty, (ii) diversity in realistic and synthesised RAG scenarios, and (iii) high utilisation in related work. For each dataset, we curate samples representative of the three types of contexts that may be encountered in realistic RAG scenarios: 1) **gold** contexts that are relevant and do not contradict LM memory, 2) **conflicting** contexts that are relevant but contradict LM memory or gold labels, and 3) **irrelevant** contexts that should be ignored by the LM (Fang et al., 2024).

A total of seven different CMTs are benchmarked on CUB, all of which are state-of-the-art representatives from the main categories of CMTs (fine-tuning, prompting techniques, mechanistic interventions, and decoding). Our results reveal several interesting findings related to the context utilisation of LMs and CMTs; context utilisation is impacted by model size, improving as the model grows in parameter count on the realistic NQ and DRUID datasets. We also find that all evaluated CMTs struggle to improve context utilisation across all context types; a CMT may be good at e.g. improving the utilisation of edited contexts but degrades performance on irrelevant contexts etc.

Taken together, Paper 5 provides insights related to context utilisation and methods for improving it. As a result, our understanding of how LMs interact with external knowledge representations is improved.

Contributions L. Hagström collected the datasets used in CUB, and implemented the prompting and PH3 CMTs. She also contributed to the design of the benchmark and made major contributions to the writing of the paper.

Y. Kim implemented the decoding and multi-agent CMTs. She also contributed to the design of the benchmark and made major contributions to the writing of the paper.

H. Yu implemented the fine-tuning CMT. She also consulted on the design of the benchmark and the writing of the paper.

S. Lee 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.

H. Cho provided supervision on the work and writing of the paper. He also contributed to the writing of the paper.

I. Augenstein provided supervision on the work and writing of the paper. She also made major contributions to the writing of the paper.

Chapter 6

Conclusions and future work

The work in this thesis has ultimately focused on knowledge and representations of information for NLP. For knowledge-intensive situations, it is important that our NLP systems can effectively represent and leverage information (RQ1). To better understand the systems and whether they can be considered reliable for knowledge-intensive tasks, it is also important that we can measure their knowledge and how information is transferred (RQ2). In this concluding chapter, we summarise the findings corresponding to our overarching research questions and consider future work. We can broadly conclude that representations of information have an important role for NLP, and that they likely will see an undiminished focus in future research.

6.1 RQ1: What are effective system designs for knowledge-intensive tasks?

Paper 1 shows that LMs are capable of acquiring parametric knowledge related to visual concepts from multimodal training. This indicates that training on visual data can be used to complement and enhance the parametric knowledge of LMs, to potentially mitigate issues stemming from the limited and biased information available in text. However, this has only been verified on simplistic knowledge-intensive tasks related to memory colours, further work is necessary to verify the potential of multimodal training for more complex knowledge-intensive tasks. Moreover, a cost analysis is also necessary to evaluate the benefits of visual training against the costs of the more complex vision-and-language models and the high computational costs associated with training on visual data.

Paper 4 similarly considers the parametric knowledge of LMs, but from a perspective of stability and interpretability. Here we find that the source and transfer of parametric knowledge is sensitive to what type of knowledge is being queried for and how well represented it is in the LM. For example, the flow of information in the LM is very different in exact fact recall scenarios compared to guesswork scenarios. The latter scenario also seems to trigger some form of parametric knowledge, but with different characteristics compared to for the former scenario. These insights bring us closer to understanding and preventing instability in LMs for knowledge-intensive tasks.

Related to effective system designs for knowledge-intensive tasks, we find in Paper 2 that RAG is more effective for improved stability compared to upscaling. This confirms prior expectations on suitable designs for improved stability and reliability. It also shows us how external information may be preferred over parametric knowledge with respect to stability. Meanwhile, Paper 5 shows how RAG benefits from upscaling, since context utilisation improves with model size. Seemingly, retrieval-augmentation together with upscaling combines the stability from retrieval-augmentation with the greater processing capacity of larger LMs, resulting in more reliable models. At the same time, we find that methods for improving context utilisation, excluding upscaling, typically work well for only one context type, while realistic RAG scenarios involve multiple.

For future work, to acquire new insights on effective systems for knowledge-intensive tasks, we may combine interpretations of the parametric knowledge of LMs with investigations into context utilisation. For example, successful context utilisation can be considered to rely on two components: 1) comprehending the provided context and 2) leveraging the provided context. Both potentially necessitate adequate parametric knowledge. It would also be interesting to expand the investigations to tool-use, another RAG-like system. Lastly, RAG promises improved interpretability by the virtue of prediction provenance, while it has yet to be investigated whether retrieval-augmented LMs are more interpretable compared to standard LMs – this would also be interesting to investigate.

To summarise the findings with respect to RQ1, this work has mainly established the necessary components of effective system designs for knowledge-intensive tasks, and overarching design recommendations for these, while more fine-grained recommendations remain to be explored. We can conclude that external information sources, accessed via retrieval-augmentation, are important for more stable and reliable systems for knowledge-intensive tasks. Seemingly, the parametric memory is a too unstable source of information compared to external non-parametric sources of information. At the same time, larger LMs are more effective at leveraging external information, making them more suitable for RAG. However, it is still not clear what makes these models better at context utilisation, what the main drivers for context utilisation are, and how these drivers interact with the parametric memory.

6.2 RQ2: How should we evaluate NLP systems for knowledge-intensive situations?

Paper 1 highlights the importance of accounting for LM instability, here expressed in terms of prompt sensitivity, for knowledge-centered evaluations.

Papers 3 and 5 further show how evaluations of RAG systems are sensitive to the underlying data, and in particular to whether the data has been synthesised or sampled from real-world scenarios. Insights gained from synthetic datasets are not guaranteed to transfer to real-world datasets, and it is therefore preferable to evaluate NLP systems on real-world data. Furthermore, Paper 4 shows how evaluations also should account for the interaction between the knowledge queried for and how well-represented it is in the model, as this is found to have a direct impact on how the query is processed by the LM.

For future work on evaluation methods for knowledge-intensive situations, we need to improve on our coverage of realistic usage scenarios. This thesis involved the development of DRUID, a dataset situated in a real-world fact-checking task. Using only DRUID, we were able to acquire several new insights related to the performance of LMs in knowledge-intensive situations. Meanwhile, there are several other interesting knowledge-intensive scenarios worthy of study. For example, it is not yet fully known how LMs behave under different knowledge-intensive domains, such as fact-checking compared to QA. Only with more evaluation datasets and corresponding evaluations can we fully map out the reliability of NLP systems for knowledge-intensive situations, and identify necessary avenues of improvement.

To summarise the findings with respect to RQ2, evaluations of NLP systems for knowledge-intensive situations should ensure that the characteristics of the evaluation dataset align with the application areas of interest and their key challenges. Insights gained from synthesised scenarios are not guaranteed to transfer to real-world scenarios, especially if the synthesised scenarios fail to represent the key characteristics of the real-world scenarios. Furthermore, evaluation methods should consider the stability and reliability of the systems under evaluation, accounting for prompt sensitivity and how different types of input may work more or less well for the system under consideration, for example in terms of irrelevant and sufficient contexts. In addition, evaluations should be careful about what types of knowledge are measured and/or inspected, and whether these should be disentangled for the evaluation.

Bibliography

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

Ronald Brachman and Hector Levesque. 2004. *Knowledge representation and reasoning*. Elsevier.

Tom Brown, 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. 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)*, pages 1860–1874, Online. Association for Computational Linguistics.

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park. 2024. Typos that broke the RAG's back: Genetic attack on RAG pipeline by simulating documents in the wild via low-level perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2826–2844, Miami, Florida, USA. Association for Computational Linguistics.

Zacharias Chrysidis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. *arXiv preprint arXiv:2404.18971*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024a. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024b. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1:1.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039, Bangkok, Thailand. Association for Computational Linguistics.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike

Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. 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*, pages 5457–5476, Singapore. Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and

Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2421–2431, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on*

Empirical Methods in Natural Language Processing: Industry Track, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*. European Conference on Computer Vision.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 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*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Joaquín Pérez-Carpinell, MD De Fez, Rosa Baldoví, and Juan Carlos Soriano. 1998. Familiar objects and memory color. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 23(6):416–427.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

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

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, Online. Association for Computational Linguistics.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. 2024. Assessing "implicit" retrieval robustness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8988–9003, Miami, Florida, USA. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023b. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*.

Jingyi Sun, Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Explaining sources of uncertainty in automated fact-checking. *Preprint*, arXiv:2505.17855.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Hugo Touvron, 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. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024a. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. *Preprint*, arXiv:2409.07394.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2024b. DAPR: A benchmark on document-aware passage retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4313–4330, Bangkok, Thailand. Association for Computational Linguistics.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of the Cognitive Science Society (CogSci)*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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)*, pages 1112–1125, Online. Association for Computational Linguistics.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, Mexico City, Mexico. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Yukun Zhu, Ryan Kiros, Rich 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 *The IEEE International Conference on Computer Vision (ICCV)*.