Understanding the Interplay between LLMs' Utilisation of Parametric and Contextual Knowledge: A keynote at ECIR 2025

Isabelle Augenstein
University of Copenhagen
Denmark
augenstein@di.ku.dk

Abstract

Language Models (LMs) acquire parametric knowledge from their training process, embedding it within their weights. The increasing scalability of LMs, however, poses significant challenges for understanding a model's inner workings and further for updating or correcting this embedded knowledge without the significant cost of retraining. Moreover, when using these language models for knowledge-intensive language understanding tasks, LMs have to integrate relevant context, mitigating their inherent weaknesses, such as incomplete or outdated knowledge. Nevertheless, studies indicate that LMs often ignore the provided context as it can be in conflict with the pre-existing LM's memory learned during pre-training. Conflicting knowledge can also already be present in the LM's parameters, termed intra-memory conflict. This underscores the importance of understanding the interplay between how a language model uses its parametric knowledge and the retrieved contextual knowledge. In this talk, I will aim to shed light on this important issue by presenting our research on evaluating the knowledge present in LMs, diagnostic tests that can reveal knowledge conflicts, as well as on understanding the characterists of successfully used contextual knowledge.

Date: 8 April 2025.

1 Introduction

LLM usage has become ubiquitous, with people using LLMs for a wide range of both creative and information-seeking tasks [Zhao et al., 2024; Chiarello et al., 2024]. In some domains such as research, information seeking is the key usage, with LLMs being used for discovering papers, generating summaries or explanations, asking broad questions about a field, or discovering topics [Liao et al., 2025].

Given the wide-ranging capabilities of LLMs, there have been suggestions of LLMs displaying Artificial General Intelligence (AGI), meaning the possession of human-like cognitive abilities. However, upon closer investigation, it becomes clear that while LLMs show high performance generally, they display several fundamental shortcomings [Bang et al., 2023]. High amounts of dataset contamination [Xu et al., 2025] mean that LLMs have been trained on the benchmark

datasets they are evaluated on already, thus not having to perform reasoning, but merely the recall of existing answers. Thus, drastic performance drops can be observed when performing small alterations to the wording used in those benchmark datasets [Yan et al., 2025; Mizrahi et al., 2024]. LLMs have also shown to perform poorly on low-resource languages [Pava et al., 2025], at most types of reasoning [Shojaee et al., 2025], and display many factual errors due to a lack of access to a knowledge base [Augenstein et al., 2024]. These issues partly arise because LLMs are developed as general-purpose models for both creative and information-seeking tasks – for the former, hallucinations might even be desirable, whereas for the latter, they are to be avoided at all costs.

There are a number of avenues that have been explored to improve the factuality of language models. Their consistency can be improved using methods including chain-of-thought prompting [Wei et al., 2022], self-consistency checking [Wang et al., 2023], continual learning [Shi et al., 2025], and knowledge editing [Wang et al., 2024a]. However, these methods all have inherent downsides. Self-consistency checking and knowledge editing is challenging as models are inherently not very consistent due to issues such as the above-mentioned prompt instability. Continual learning is by nature highly costly. Knowledge editing can lead to ripple effects, i.e. the editing of knowledge beyond what is intended [Cohen et al., 2024] as well as the removal of long-tail knowledge, due to knowledge superposition [Hu et al., 2025], and identifying what knowledge to edit is a task in and of itself [Yu et al., 2025]. Thus, internal consistency checking only partly addresses the factuality issues of LLMs. Another direction is the combination with external knowledge, by detecting and correcting factual mistakes at inference time [Wang et al., 2024b], using a modularised knowledge-grounded framework [Arakelyan et al., 2025], or, very commonly, retrieval-augmented generation (RAG) [Fan et al., 2024]. These can better take the context-dependent nature of queries into account, by retrieving *contextual knowledge* to augment the LLM's *parametric knowledge*.

However, a key research gap is that the interplay between contextual and parametric knowledge underexplored, as is when contextual knowledge even should overwrite parametric knowledge. This keynote explores these two topics by highlighting recent research conducted in our research group.

2 Determining what Parametric Knowledge influences a LLM's Prediction

Parametric knowledge is the knowledge encoded in a LM's weights which an LM acquires during training. Our study [Yu et al., 2024] focuses on the changes in knowledge acquired during LLM training and task-adaptive training for knowledge-intensive tasks, such as fact checking, QA, and natural language inference. To unveil LM's parametric knowledge used to arrive at a prediction, attribution methods are commonly used. Previous methods operate on different levels (e.g. instance vs. neuron) and are studied in isolation, with no consensus as to which methods work best best in which scenarios. We thus propose a unified evaluation framework, illustrated in Figure 1, that compares these two streams of attribution methods, to provide a comprehensive understanding of a LM's inner workings. The framework includes methods to align the most influential training instances with the most important neurons. The success of this alignment procedure is evaluated using faithfulness tests for sufficiency (i.e. the activation of key neurons) and complete-

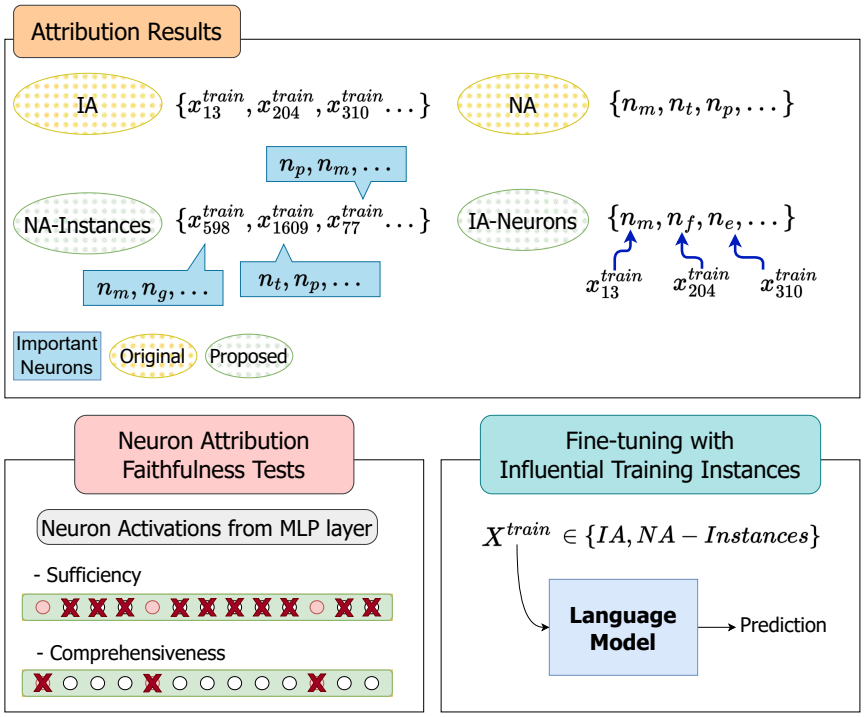

Figure 1. Evaluation framework comparing Instance and Neuron Attribution methods [Yu et al., 2024].

ness (the suppression of the activation of key neurons), along with an evaluation of the impact of fine-tuning with influential training instances.

We analyse the MLP classification layers instead of entire model weights due to the high computational cost, as prior worked shows that this layers shows a high correlation with the overall model weights [Pezeshkpour et al., 2021]. Our findings show that most MLP neurons can be removed without significant changes to the predictions compared to the original model, contradicting the assumption of prior work that most model's knowledge is located in these neurons. We hypothesise that this might be due to the importance of attention weights for encoding knowledge, as also argued by prior work [Wiegreffe and Pinter, 2019]. Our experiments on fine-tuning with influential training instances identified using Instance Attribution leads to results on par with the equivalent number of randomly selected training instances. A regression test reveals that this is likely due to the diversity of highly influential training instances being significantly lower than that of randomly selected training instances, which hampers the effectiveness in reaching a high performance with the fine-tuned model. However, we find a potential application in combining Instance and Neuron Attribution for the discovery of dataset artifacts [McCoy et al., 2019], specifically models overfitting to certain lexical patterns observed at training time, with the combined methods being more effective at discovering such instances than either Neuron or Instance Attribution alone. Overall, we find that Instance Attribution and Neuron Attribution result in different explanations about the knowledge responsible for the test prediction.

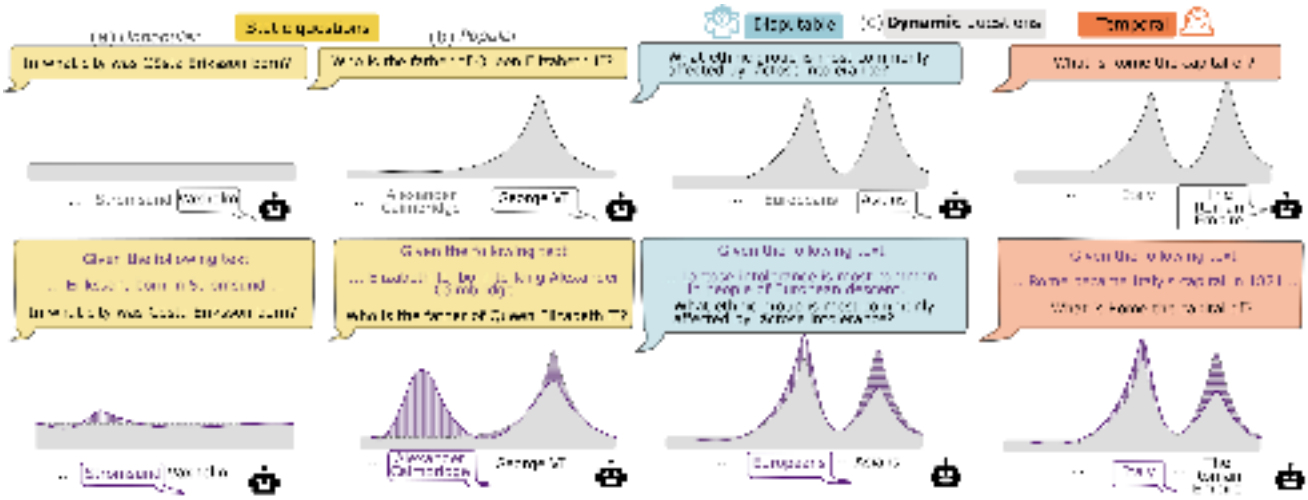

Figure 2. Instances with varying dynamicity in the DynamicQA dataset [Marjanovic et al., 2024].

3 Revealing Conflicts between Parametric and Contextual Knowledge

Though attribution methods methods, as studied in the previous section, can theoretically pinpoint the knowledge a prediction is based on, their application is expensive, only possible given access to full model parameters, and their application to only partial model parameters can result in unintuitive or contradictory findings. An alternative approach is to use probing tests, which we use to study two different types of knowledge conflicts: *intra-memory conflict*, the conflict caused by contradicting representations of the fact within the training data, can cause uncertainty and instability of an LM, and *context-memory conflict*, the conflict caused by the context contradicts to the parametric knowledge [Marjanovic et al., 2024]. We study these two phenomena in the context of question answering, where fact dynamicity can impact the knowledge that should be used. To this end, our dataset DynamicQA contains samples of *static facts*, which only have one possible representation; *temporal facts*, which change over time; and *disputable facts*, which can change depending on the viewpoint (see Figure 2).

We then study how the output distribution varies depending on these types of facts, as well as when additional context is provided, as is done in RAG settings. Dynamic facts (both temporal and disputable ones) should then lead to intra-memory conflicts, and presenting models with conflicting context during RAG should lead to context-memory conflicts. We also expect to observe differences between popular facts, for which a model is expected to be much more certain than unpopular facts. We measure intra-memory conflict using semantic entropy, which captures the semantic variation present in parametric memory [Kuhn et al., 2023], and context-memory conflict using a measure we term *coherent persuasion score*, based on Du et al. [2024] to approximate an LM's semantic shift in output distribution given competing context.

Our findings reveal that, counter-intuitively, presenting LMs with manipulated contexts of static facts and facts with low dynamicity results in greater persuasiveness for LMs than dynamic facts – or, put differently, facts that change regularly are less likely to be updated with context-retrieval, yet facts that never change are easily persuaded. This fact dynamicity is found to be the

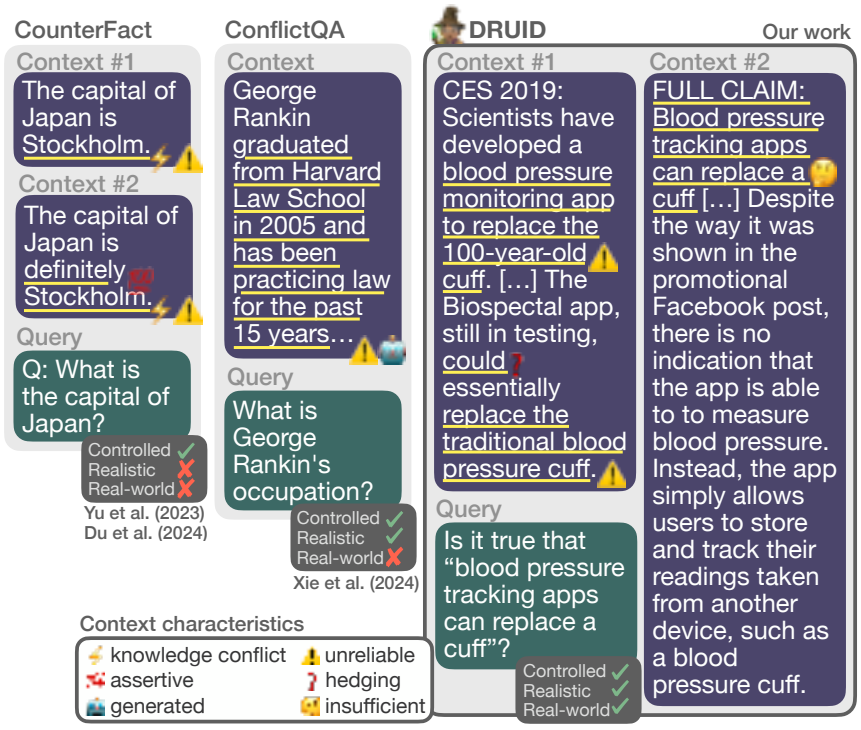

Figure 3. The DRUID dataset [Hagström et al., 2025] in comparison with prior synthetic datasets.

strongest, most consistent negative indicator of model persuasion across models, outperforming fact popularity, which was previous used to guide RAG models. These results underline the need for new measures of intra-memory conflict and the need for other indicators of successful context usage in RAG, especially for low-certainty domains [Ni et al., 2024].

4 Determining when or how RAG uses Contextual Knowledge

Successful RAG not only relies on the usage of retrieved information, but also on the retrieval of relevant information, and the interplay between these two components, though prior work studies these aspects in isolation. As such, little is understood about the characteristics of retrieved content, and its impact on LLM usage. Though some context usage studies exist, they use synthetic data, and thus do not reflect real-world RAG scenarios. To overcome this, we propose new dataset for claim verification with real-world contexts to measure realistic context usage (DRUID), a novel context usage measure (ACU), and generate novel insights into LLMs' context usage characteristics [Hagström et al., 2025].

Some patterns we find are that context from fact-check sources leads to high context usage, which is likely due to the higher rate of assertive and to-the-point language. Also, the more direct discussion of claims with multiple arguments might make it more more convincing to the LM. Similarly, evidence documents published after the claim and gold evidence sources lead to higher context usage. Conversely, references to external sources show low correlations with ACU.

Moreover, LLMs prioritise contexts with high query-context similarity, which are more difficult to obtain in real-world RAG setting. Lastly, LMs are shown to be less faithful to long contexts.

Comparing DRUID to synthetic datasets, we find that synthetic datasets oversell the impact of certain context characteristics (e.g. knowledge conflicts), which are rare in retrieved data. Synthetic data exaggerates context repulsion, which is rarer for realistic data. While we identify different characteristics indicating RAG failure in real-world settings, there is no single indicative characteristic. This provides a reality check on LLM context usage, and underscores the need for real-world aligned studies to understand and improve context use for RAG.

In some follow-up work, we study richer forms of interaction between parametric and contextual knowledge for RAG, namely complementary or supporting knowledge [Islam et al., 2025]. There, we find that for claim verification, knowledge conflicts often pushes models to prefer contextual rather than parametric knowledge, whereas for common-sense question answering with fewer knowledge conflicts, LLMs more often rely on their parametric memory.

Moreover, we perform a study on how context-memory conflicts can be resolved using a wide range of so-called *context manipulation techniques (CMT)* [Hagström et al., 2025], going beyond simply providing the context as done in Marjanovic et al. [2024]; Hagström et al. [2025]. These range from simple methods such as prompting to more elaborate methods including mechanistic interventions. The study is performed on the datasets as Hagström et al. [2025]. We find that, overall, there is no clear winner between the different CMTs. Moreover, larger LLMs on average perform better than smaller LLMs at successfully using context, though smaller models can outperform them with the best CMT for the specific model.

5 Concluding Remarks

Despite ongoing work, the inner workings of RAG-based language models, and how and when they use or ignore context is still poorly understood. This is a unique challenge necessitating collaborations between researchers working on information retrieval and natural language processing. This keynote has aimed to shed light on some aspects of this topic, though often, and notwithstanding rigorous evaluation efforts, our research has unearthed more questions than it has provided answers.

Some effects we observed have been reminiscent of trends observed for earlier neural methods used in NLP, namely pre-trained language models. These include that LLMs are excellent at recitation, not at reasoning [Yan et al., 2025], which could also be observed for PLMs [Petroni et al., 2019]. The fact that RAG-based claim verification models prioritise easy-to-understand sources [Hagström et al., 2025] can also be traced back to effects observed for PLMs [Augenstein et al., 2019].

As this ECIR 2025 keynote was a Karen Spärk Jones award lecture, I would like conclude by honouring her memory in sharing one of her famous quotes:

Those [...] who had been around for a long time, can see old ideas reappearing in new guises [...]. But the new costumes are better made, of better materials, as well as more becoming: so research is not so much going round in circles as ascending a spiral (Karen Spärk Jones, 1994).

Acknowledgements

I would like to thank The Chartered Institute for IT (BCS) as well as Bloomberg for honouring me with the Karen Spärck Jones Award, as well as my research group, without whom none of this would have been possible.

This research was in large parts supported by the European Union (ERC, ExplainYourself, 101077481). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Erik Arakelyan, Pasquale Minervini, Patrick Lewis, Pat Verga, and Isabelle Augenstein. FLARE: Faithful Logic-Aided Reasoning and Exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23396–23414, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1193. URL https://aclanthology.org/2025.emnlp-main.1193/.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 4685–4697, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1475. URL https://aclanthology.org/D19-1475/.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.*, 6(8):852–863, 2024. URL https://www.nature.com/articles/s42256-024-00881-z.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.45. URL https://aclanthology.org/2023.ijcnlp-main.45/.

Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, and Gualtiero Fantoni. Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation*, 133:103002, 2024. ISSN 0166-4972. doi: https://doi.org/10.1016/j.technovation.2024.103002. URL https://www.sciencedirect.com/science/article/pii/S016649722400052X.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. Context versus Prior Knowledge in Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.714. URL https://aclanthology.org/2024.acl-long.714.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19691–19730, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.968. URL https://aclanthology.org/2025.acl-long.968/.

Lovisa Hagström, Youna Kim, Haeun Yu, Sang-Goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. CUB: Benchmarking Context Utilisation Techniques for Language Models, 2025. URL https://arxiv.org/abs/2505.16518.

Chenhui Hu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge in Superposition: Unveiling the Failures of Lifelong Knowledge Editing for Large Language Models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i22.34583. URL https://doi.org/10.1609/aaai.v39i22.34583.

Sekh Mainul Islam, Pepa Atanasova, and Isabelle Augenstein. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement, 2025. URL https://arxiv.org/abs/2511.01706.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. In *COLM*, 2025. URL https://openreview.net/forum?id=p0BwJk3R1p.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.838. URL https://aclanthology.org/2024.findings-emnlp.838/.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681. URL https://aclanthology.org/2024.tacl-1.52/.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 11375–11388, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.675. URL https://aclanthology.org/2024.findings-acl.675.

Juan Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts, 2025. URL https://hai.stanford.edu/assets/files/hai-taf-pretoria-white-paper-mind-the-language-gap.pdf.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, and Sameer Singh. An Empirical Comparison of Instance Attribution Methods for NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 967–975, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.75. URL https://aclanthology.org/2021.naacl-main.75/.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual Learning of Large Language Models: A Comprehensive Survey. *ACM Comput. Surv.*, 58(5), November 2025. ISSN 0360-0300. doi: 10.1145/3735633. URL https://doi.org/10.1145/3735633.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, 2025. URL https://arxiv.org/abs/2506.06941.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge Editing for Large Language Models: A Survey. *ACM Comput. Surv.*, 57(3), November 2024a. ISSN 0360-0300. doi: 10.1145/3698590. URL https://doi.org/10.1145/3698590.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *ICLR*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.830. URL https://aclanthology.org/2024.findings-emnlp.830/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Infini-gram mini: Exact n-gram Search at the Internet Scale with FM-Index. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24955–24980, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1268. URL https://aclanthology.org/2025.emnlp-main.1268/.

Kai Yan, Yufei Xu, Zhengyin Du, Xuesong Yao, Zheyu Wang, Xiaowen Guo, and Jiecao Chen. Recitation over Reasoning: How Cutting-Edge Language Models Can Fail on Elementary School-Level Reasoning Problems? *CoRR*, abs/2504.00509, April 2025. URL https://arxiv.org/abs/2504.00509.

Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8186, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.444. URL https://aclanthology.org/2024.acl-long.444/.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. Entangled in Representations: Mechanistic Investigation of Cultural Biases in Large Language Models. *CoRR*, abs/2508.08879, August 2025. URL http://dblp.uni-trier.de/db/journals/corr/corr2508.html#abs-2508-08879.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *ICLR*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.