Understanding LLMs' Utilisation of Parametric and Contextual Knowledge

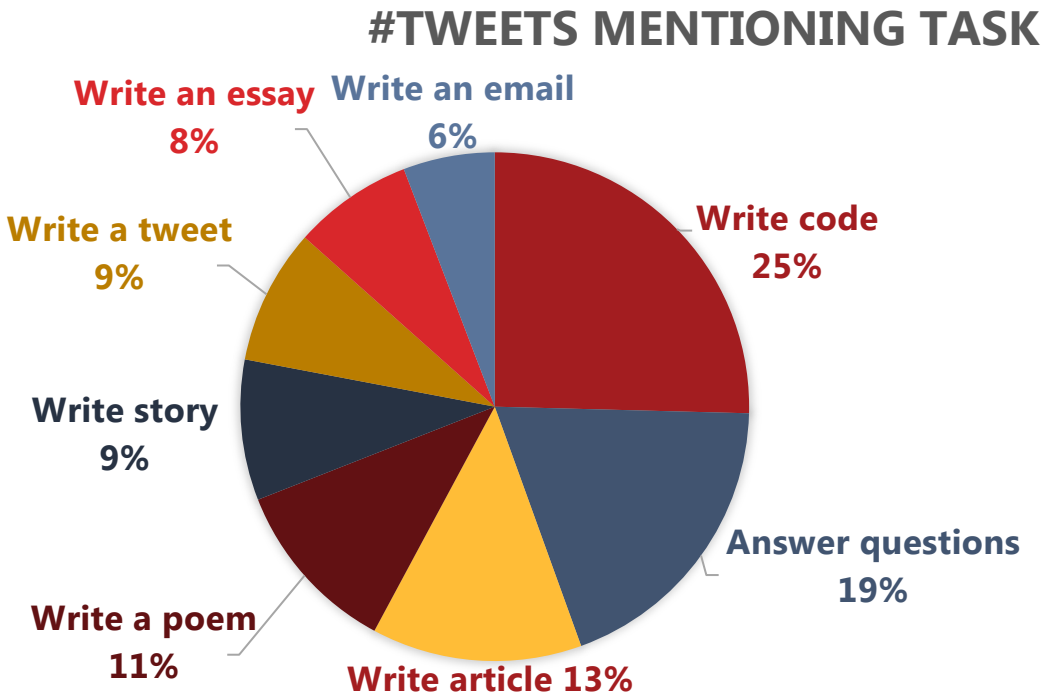
Isabelle Augenstein

EurIPS 2025 Workshop on "The Science of Benchmarking and Evaluating AI"
6 December 2025

Funded by the European Union
European Research Council
Established by the European Commission

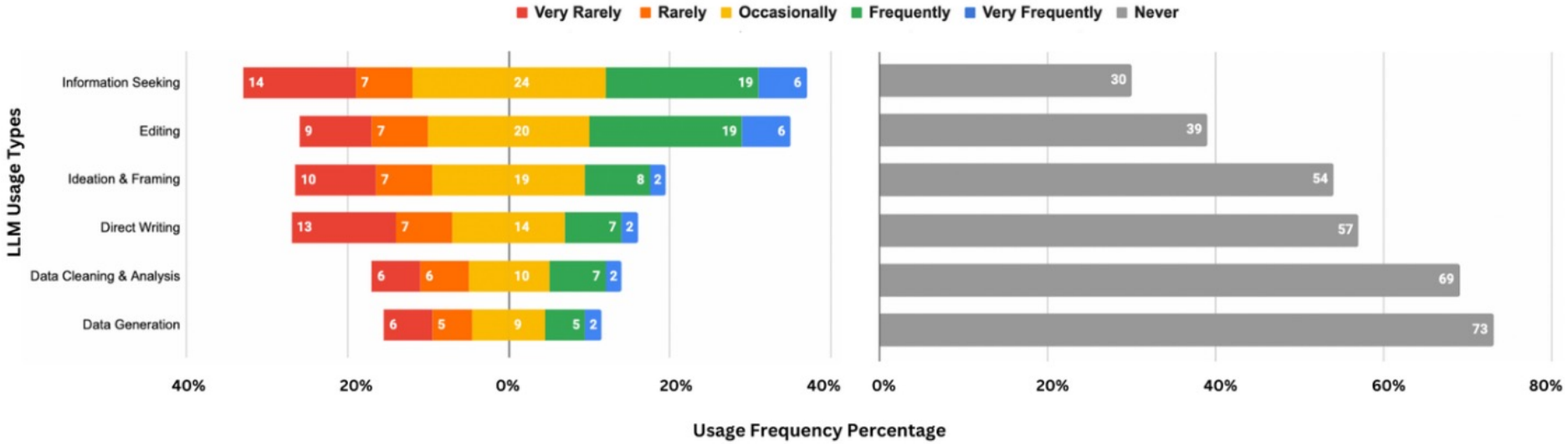
UNIVERSITY OF COPENHAGEN

Usage of Large Language Models



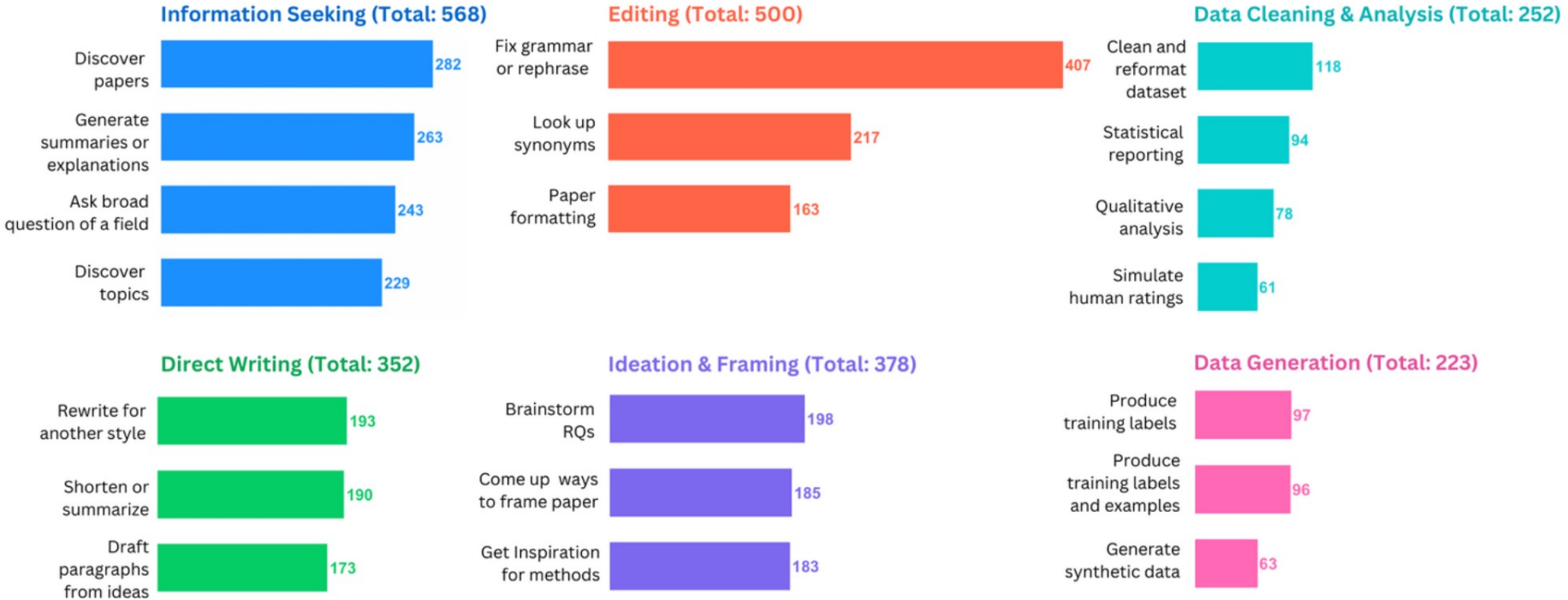
Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, Gualtiero Fantoni. Future applications of generative large language models: A data-driven case study on ChatGPT. Technovation Volume 133, May 2024, 103002.

Usage of Large Language Models



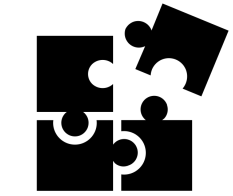
Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. ArXiv, abs/2411.05025.

Usage of Large Language Models



Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. ArXiv, abs/2411.05025.

Factuality Challenges of Large Language Models

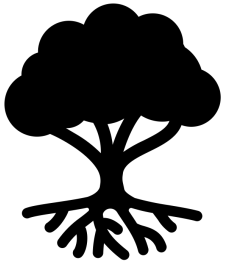


Citation Gaps

Truthfulness

Fluent Style

Outdated Knowledge



Grounding Deficiency

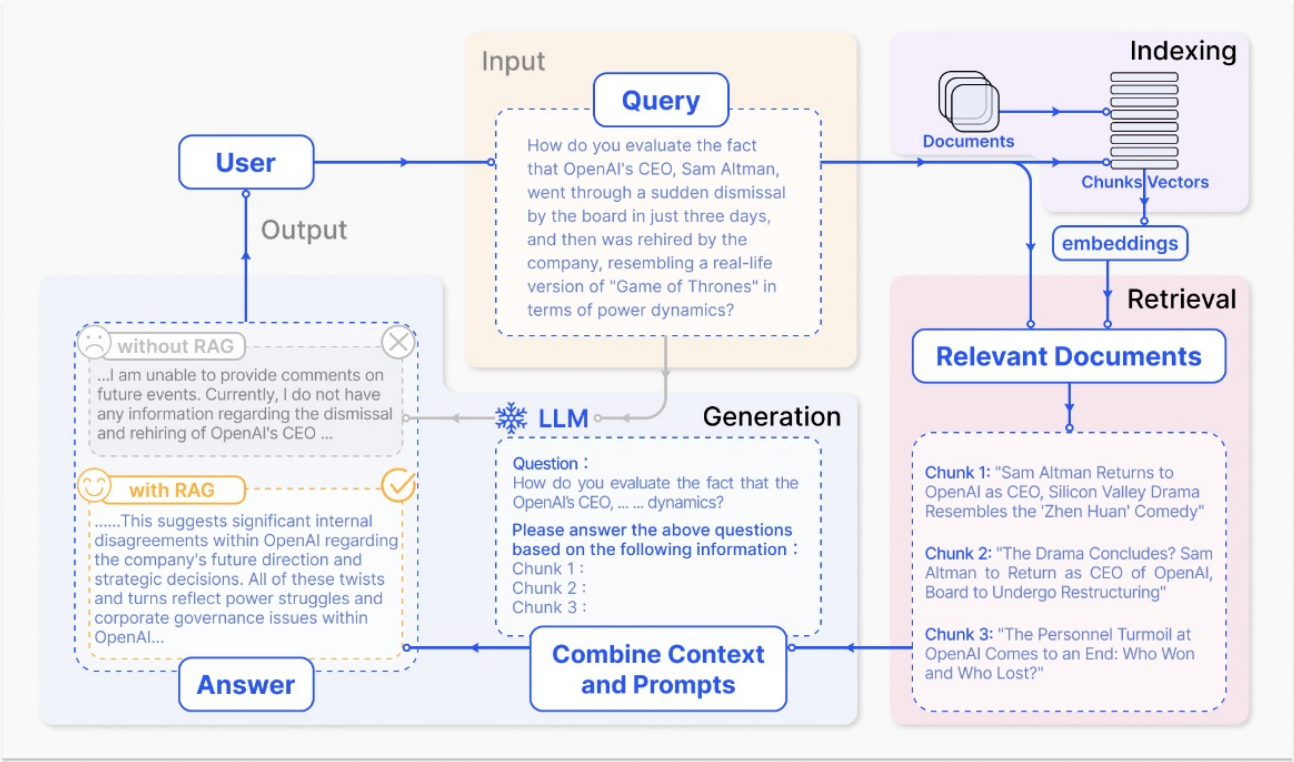
Confident Tone

Halo Effect

Unreliable Evaluation

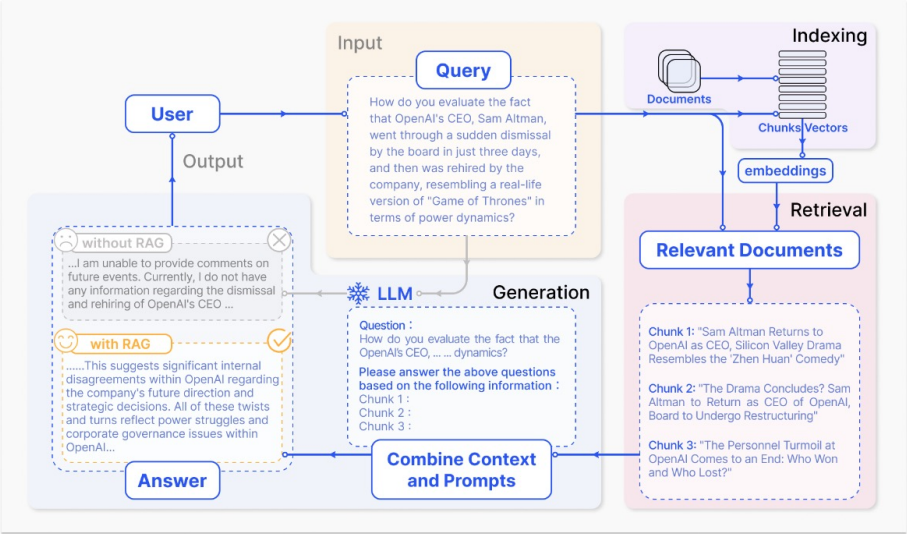
Augenstein et al. (2024). Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.

Augmentation of LLMs with External Knowledge



Gao et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arxiv:2312.10997.

Augmentation of LLMs with External Knowledge

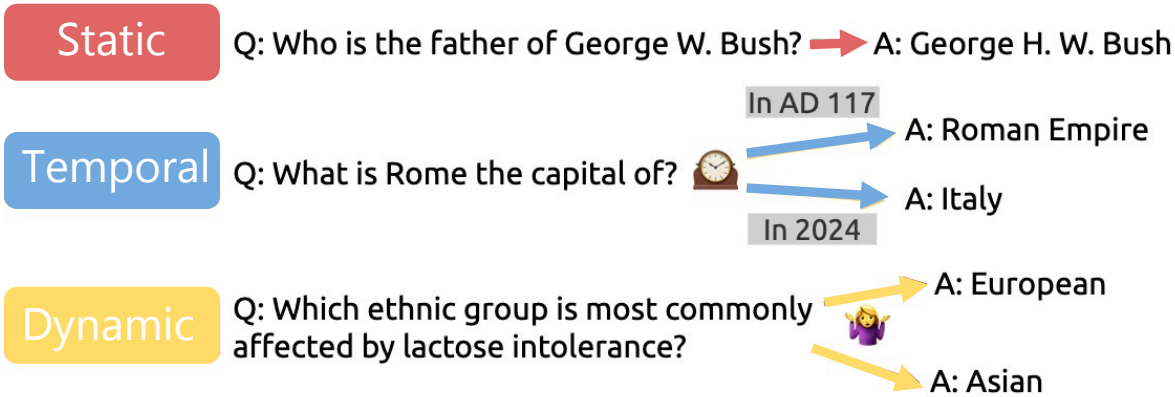


- *Retrieving contextual knowledge to augment LLM's parametric knowledge*
- *Can better take context-dependent nature of queries into account*
- *Interplay between contextual and parametric knowledge underexplored*
- *When should contextual knowledge overwrite or augment parametric knowledge?*

Overview: Understanding LLMs' Knowledge Utilisation

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

Fact Dynamicity and Knowledge Conflicts



- Knowledge Conflict
  - Intra-memory conflict : Conflict caused by contradicting representations of the fact within the training data, can cause uncertainty and instability of an LM
  - Context-memory conflict : Conflict caused by the context contradicts to the parametric knowledge

We investigate the impact of fact dynamicity on LLM output in question answering

Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

DynamicQA

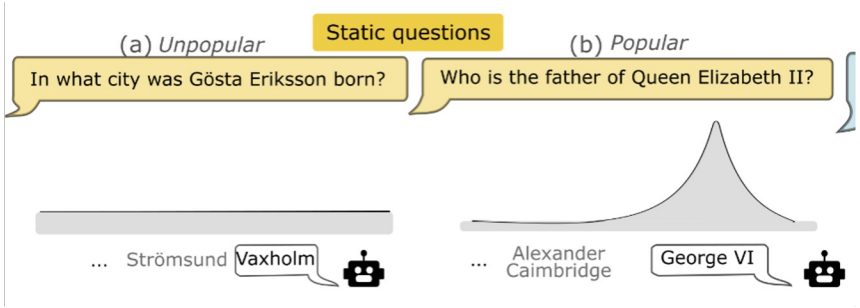
We release a dataset of 11,378 questions and answers.

- We identify **temporal** relations as relations with >1 edit on Wikidata
- We identify **static** relations as relations with no edits on Wikidata
- We identify **disputable** relations as sentences with >1 *mutual reversions* on Wikipedia (*Controversial topics*)

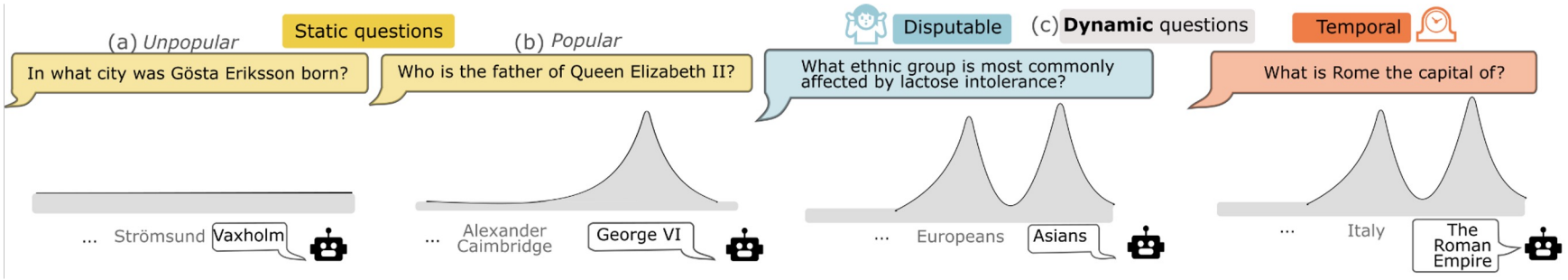
For each relation, we use the edited object as the **answer** and formulate a **question.**

We retrieve relevant **context** mentioning the subject and object from *Wikipedia*.

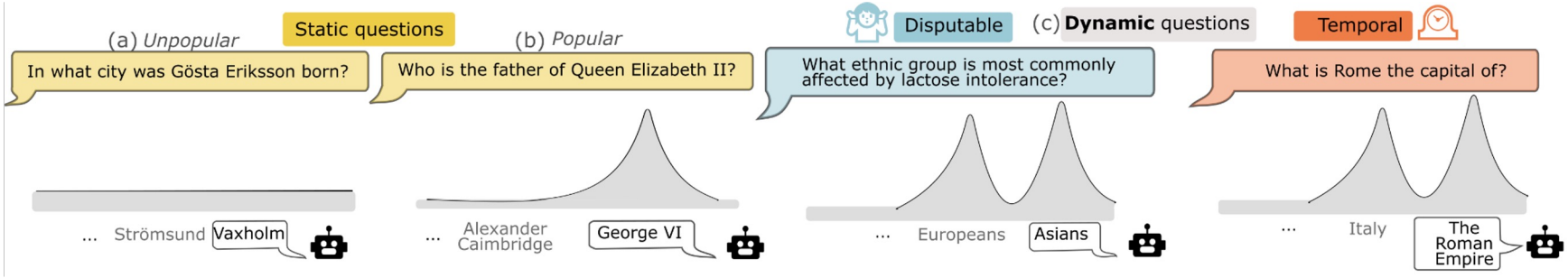
Intra-Memory Conflict in Output Distribution



Intra-Memory Conflict in Output Distribution



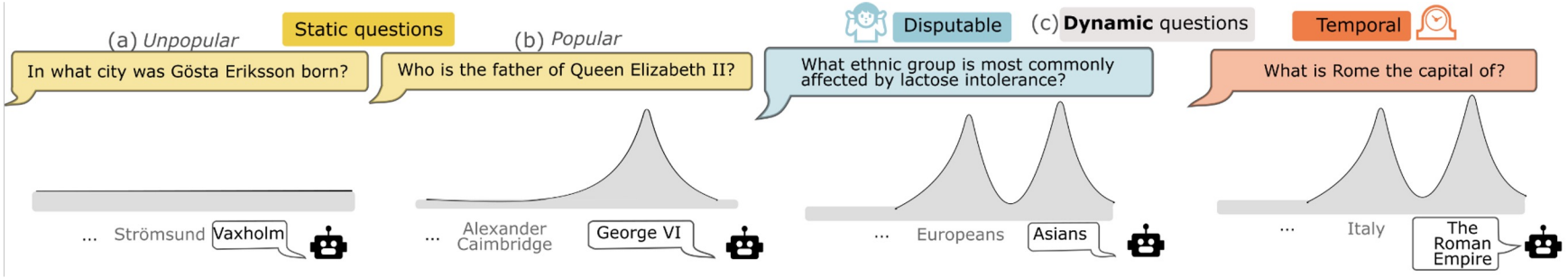
Intra-Memory Conflict in Output Distribution



***Dynamic* facts should show greater *entropy* across objects.**

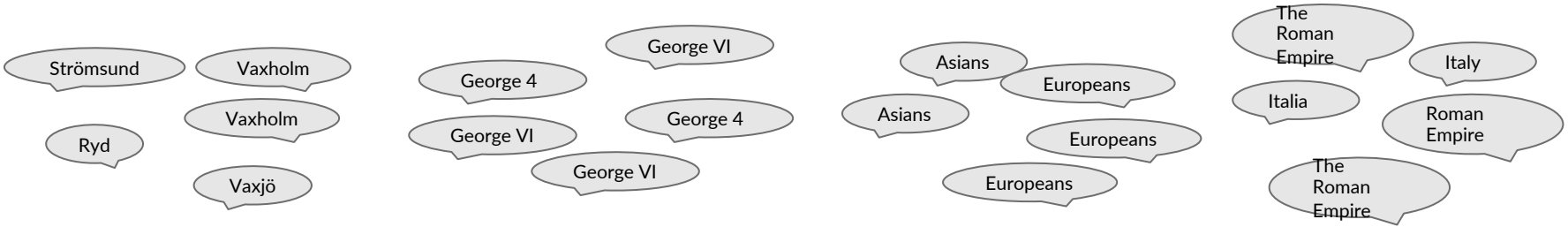
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*

Intra-Memory Conflict in Output Distribution

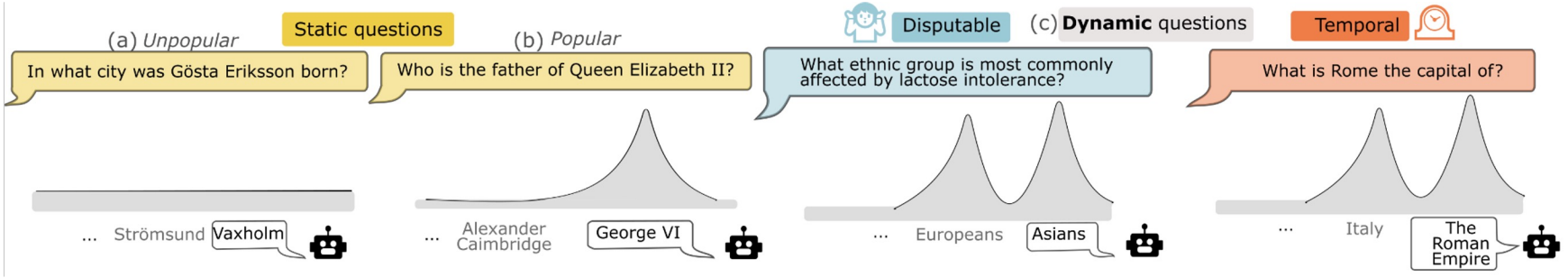


***Dynamic* facts should show greater *entropy* across objects.**

We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*

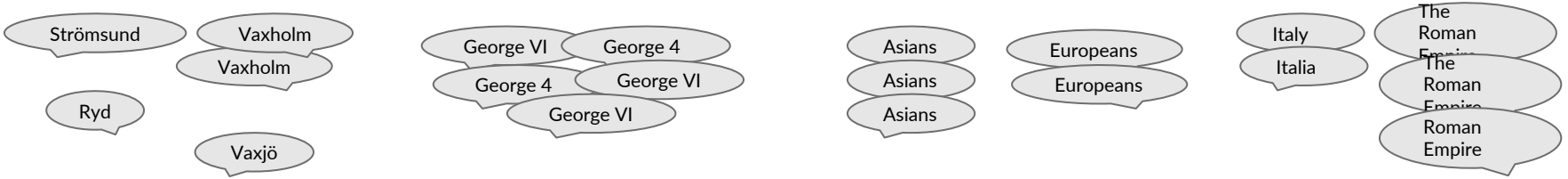


Intra-Memory Conflict in Output Distribution

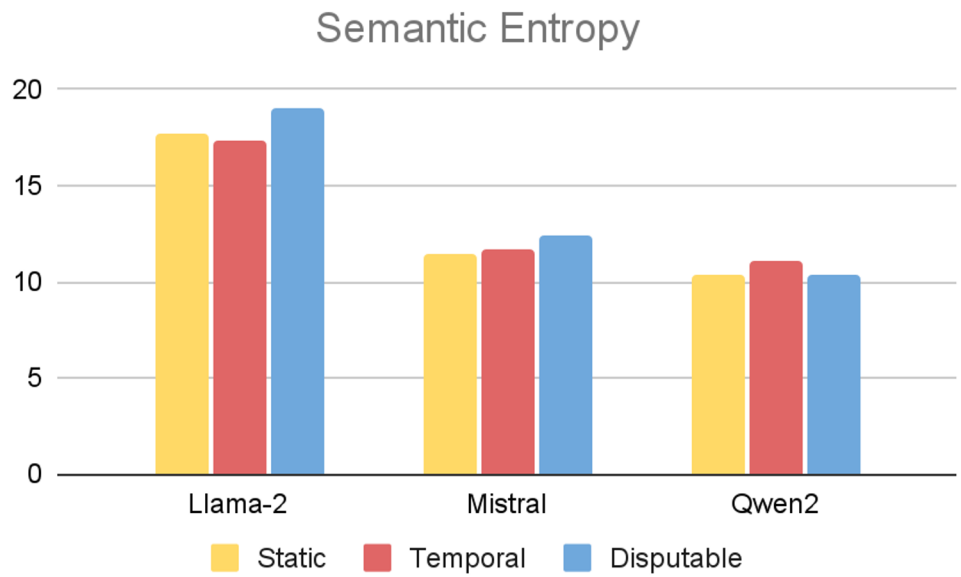


***Dynamic* facts should show greater *entropy* across objects.**

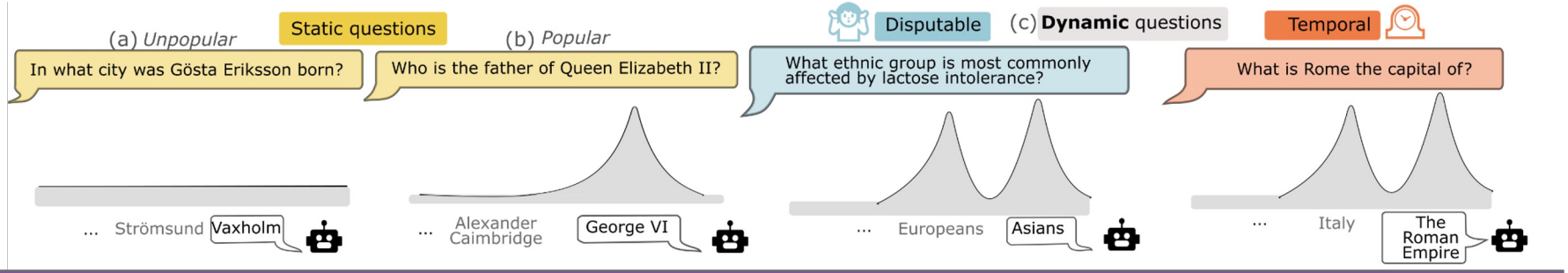
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*



However, this is not always the case

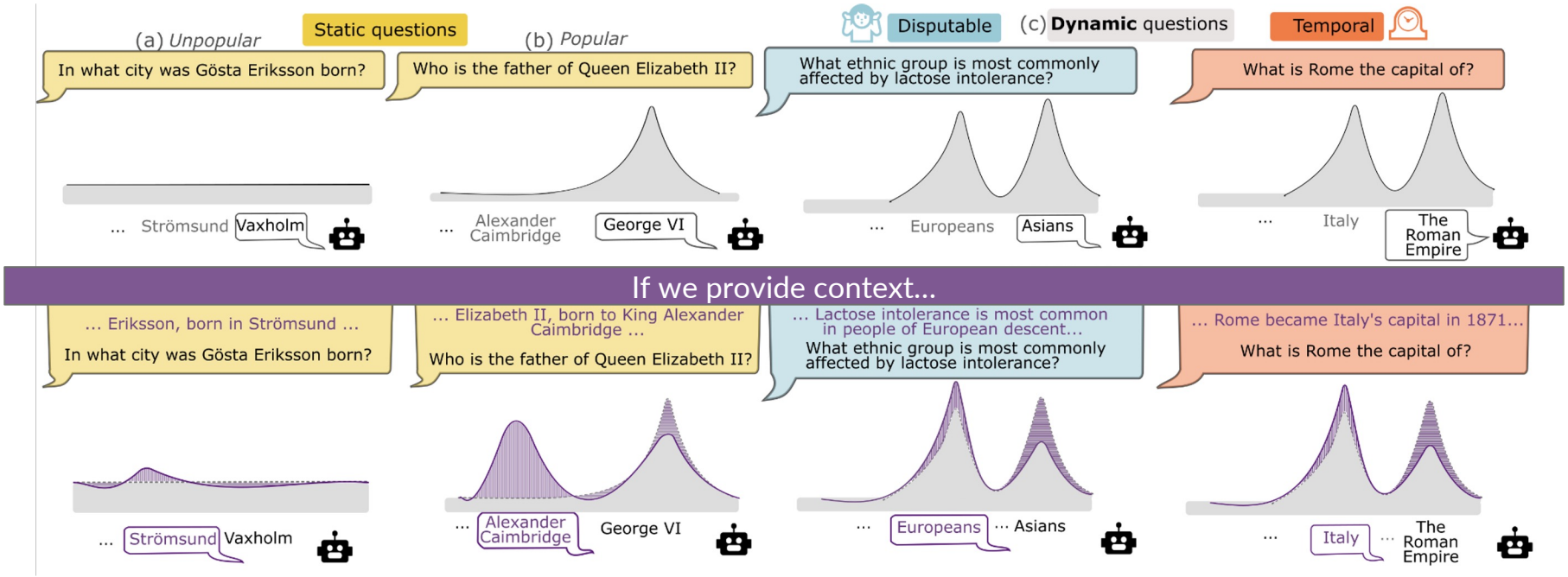


Context-Memory Conflict

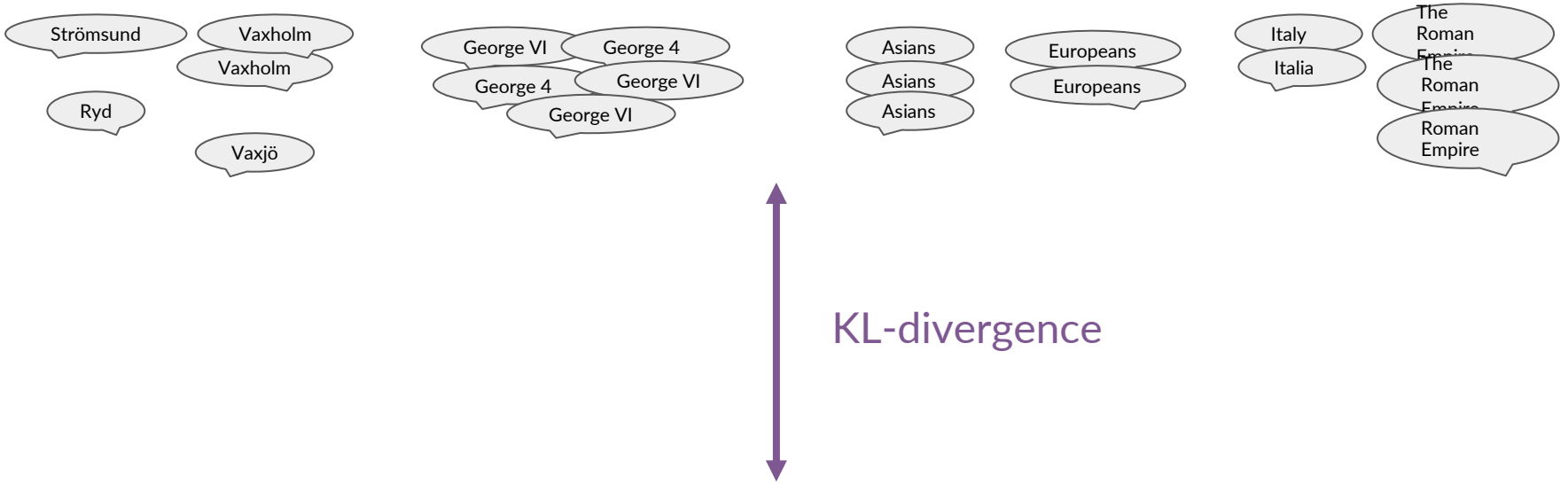


If we provide context...

Context-Memory Conflict

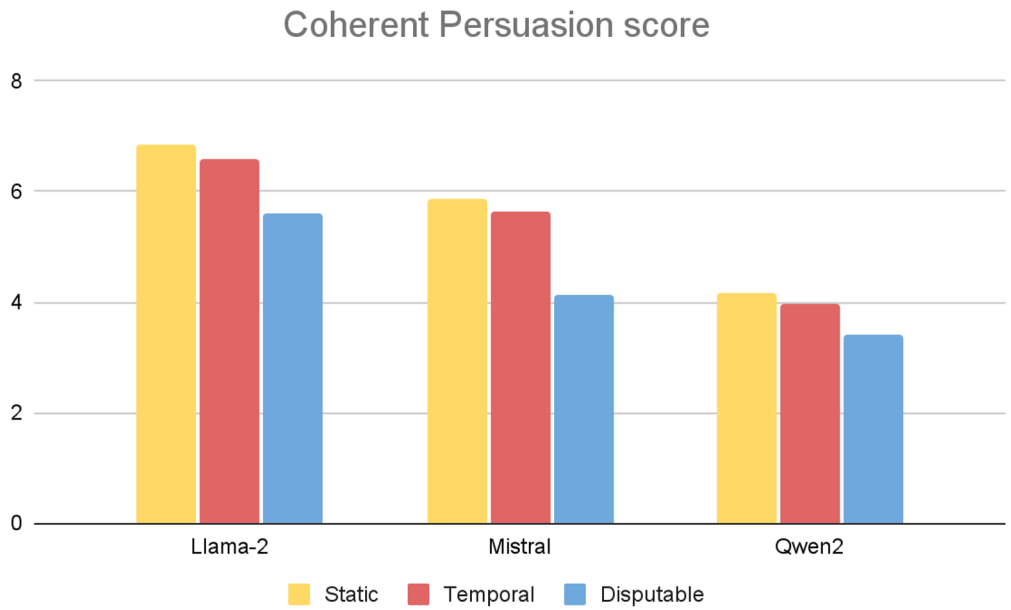


Coherent Persuasion Score



Persuasion Score across Partitions

We see the **greatest persuasion score** for the **static dataset**.



Persuasion Score across Partitions

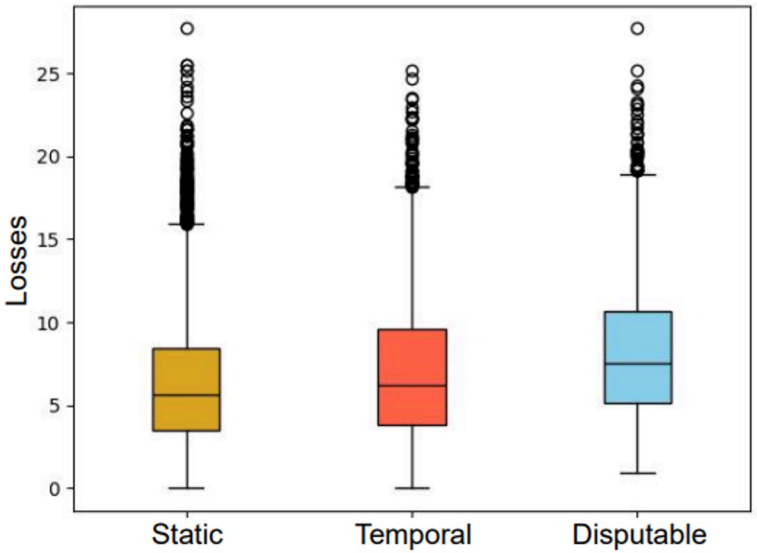
We see the **greatest persuasion score** for the **static dataset.**

However, this is **successful persuasion**, in that the model output distribution has been changed.

How far are we from from successful persuasion for dynamic facts?

→ *Loss (target answer | question) (~ Perplexity)*

Loss across Partitions



Loss reflects the likelihood of an output given the model's trained parameters.

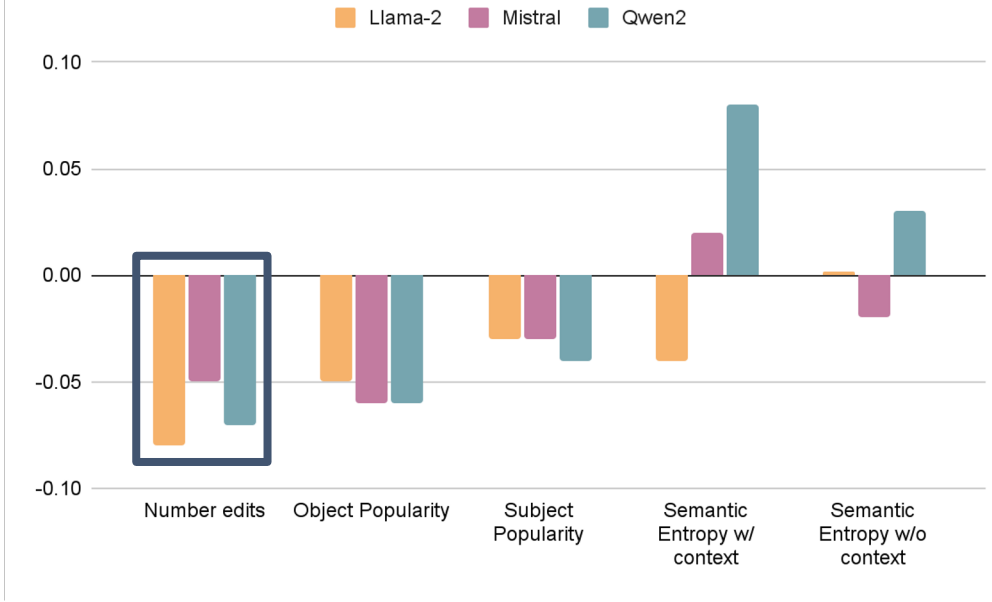
A higher loss indicates greater change required to steer the LM to output the target answer.

It requires more change in the model's parameters to obtain the desired answer for **temporal** and **dynamic** facts ($p<<<10^{-5}$).

This **cannot** be accomplished by **context alone.**

What impacts Persuasion? Predictors of Persuasion

Logistic regression model to predict if an instance will be **stubborn** or **persuaded**



Number of edits is the **strongest**,

most consistent negative indicator of model persuasion across models

Implications: Knowledge Conflict and Fact Dynamicity

- **Temporal and disputable facts**, which have greater historical variability (which is expected to be reflected in a training dataset, leading to intra-memory conflict):
  - Show lower persuasion scores, fewer persuaded instances, more stubborn instances
  - ➢ Are less likely to be updated with context, instead requiring models to be retrained or manually edited to reflect changing information.
- **Fact dynamicity (number of edits)** has a greater impact on a model's likelihood for persuasion than a fact's popularity
  - Fact popularity often used to guide RAG in previous literature
  - ➢ Other approaches might be required for retrieval augmentation in low-certainty domains

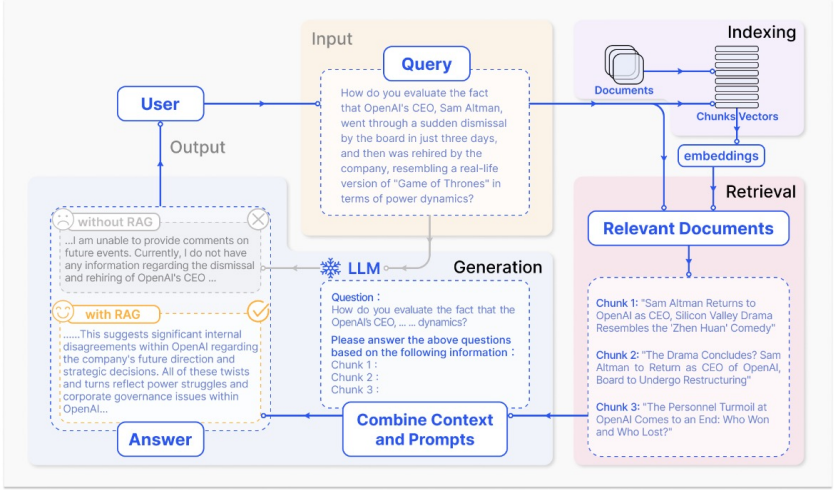
Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Overview: Understanding LLMs' Knowledge Utilisation

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

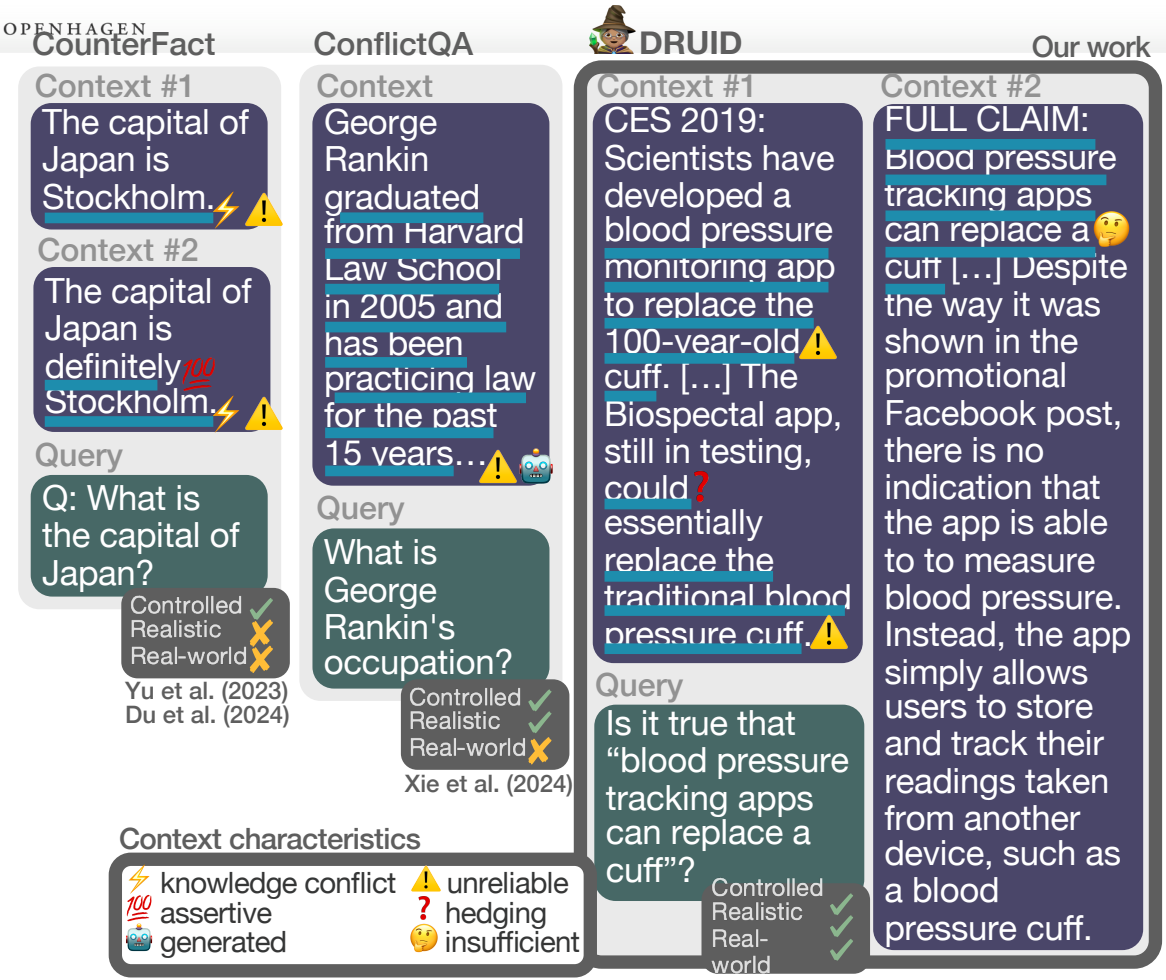
Context Utilisation of Retrieval-Augmented Generation

- Successful RAG requires
  - Retrieval of relevant information
  - Successful use of retrieved information by LLM
- Prior work studies these aspects in isolation
  - Little understood about characteristics of retrieved content; and impact on LLM usage
  - Context usage studies use synthetic data
  - Do not reflect real-world RAG scenarios



Contributions:
- new dataset to measure realistic context usage (DRUID)
- novel context usage measure (ACU)
- insights into LLMs' context usage characteristics

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

DRUID data selection process

- Crawl 7 geographically diverse English language fact checking datasets for claims
  - Collapse labels
- Retrieve relevant evidence pages
  - 20 from Google Search, 20 from Bing Search
  - De-duplicate results

Source	#claims	#samples	IAA
checkyourfact	220	890	0.77
science.feedback	220	913	0.64
factcheckni.org	109	429	0.50
factly	180	739	0.80
politifact	220	931	0.74
srilanka.factcrescendo	156	598	0.75
borderlines	224	990	0.53
Total	1,329	5,490	0.71

Our label	Incoming label
True	True
	TRUE
	ACCURATE
	ACCURATE WITH CONSIDERATION
	Correct
	Mostly accurate
	Accurate
Half-true	Half True
	PARTLY TRUE
	Correct But...
	Mostly_Accurate
	Partially correct
False	False
	FALSE
	MISLEADING
	Misleading
	Inaccurate
	Incorrect, Flawed_Reasoning
	INACCURATE
	INACCURATE WITH CONSIDERATION

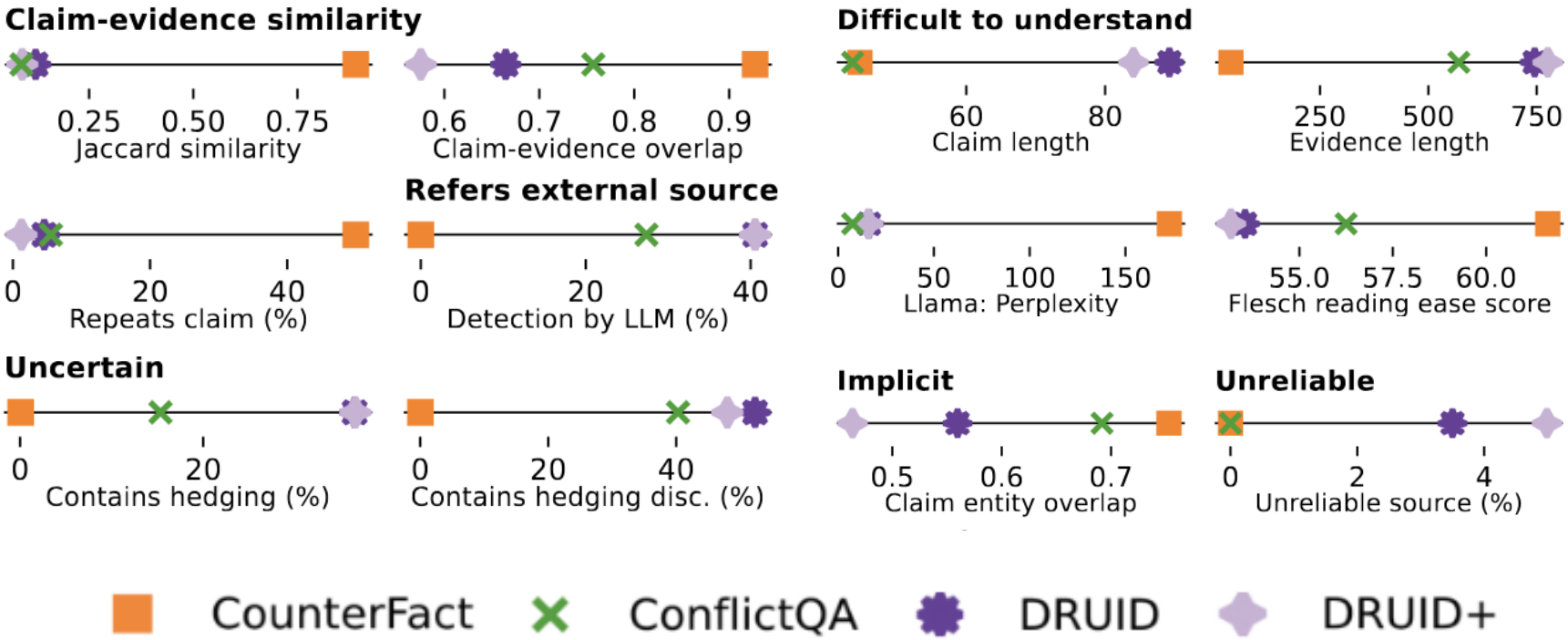
Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

DRUID content characteristics

- **Context-memory conflicts less prevalent in real-world scenarios**
- Measured as share of samples for which the stance of the provided evidence conflicts with the parametric model prediction (no context or evidence provided)
- For Llama 3.1 8B, e.g.:
  - CounterFact: 97.41% of supporting evidence
  - ConflictQA: 71.16% of refuting evidence
  - DRUID: 58.09% of supporting evidence
- Overall, rates of memory conflicts sizably lower for DRUID than for synthetic datasets

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

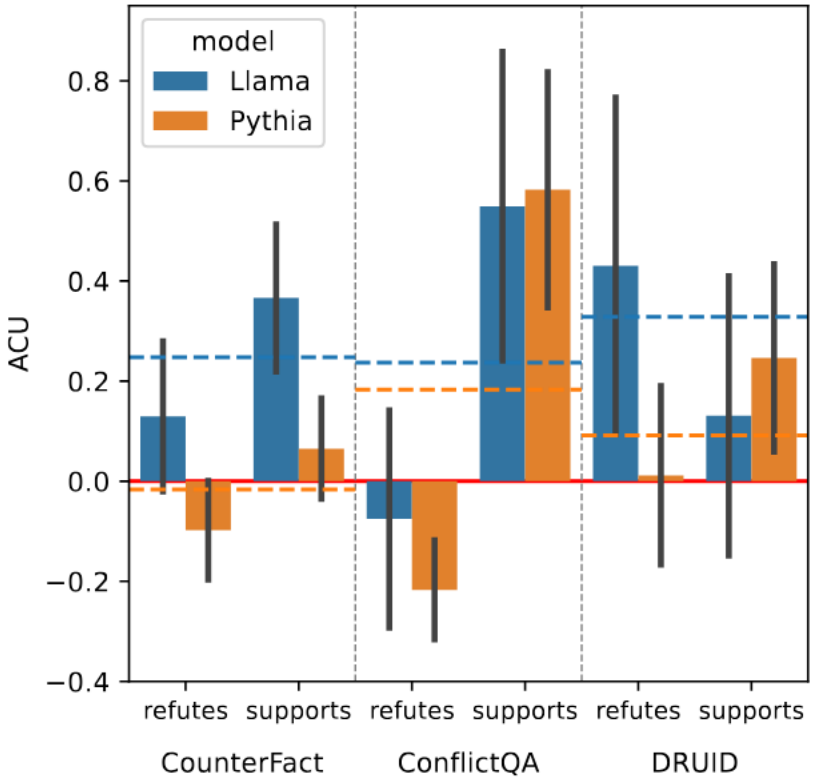
DRUID content characteristics ctd



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Context utilisation of RAG

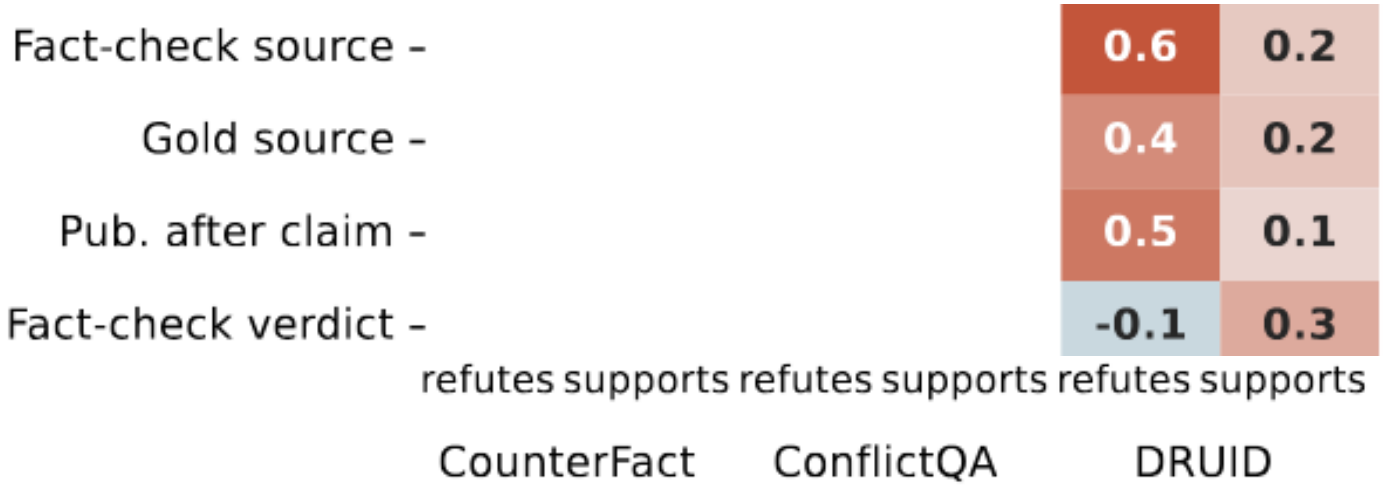
- Context usage (ACU score):
  - Re-scaled difference in salient token probability for different labels for a claim between settings with vs. without evidence
- **Synthetic** datasets:
  - Over-prefer supporting evidence
  - Context repulsion for refuting evidence
  - Generated automatically -> aligned with parametric memory
- Real-world dataset:
  - Context utlisation and repulsion both lower



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

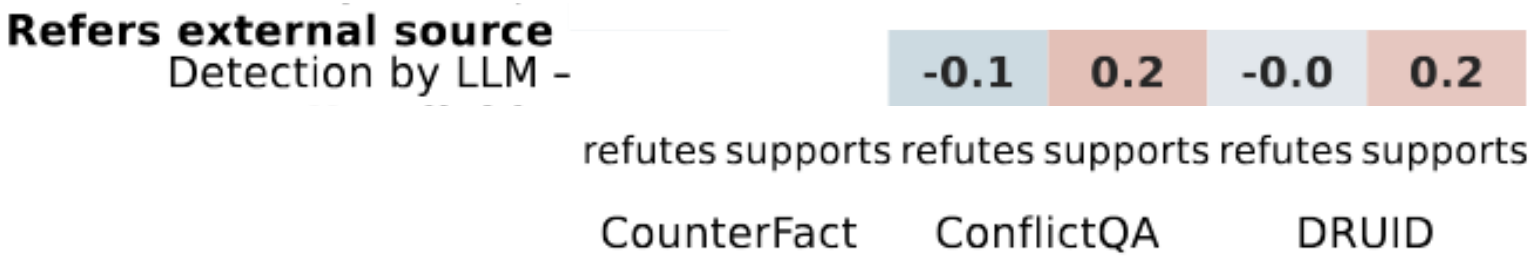
Influence of content characteristics on RAG

- **Context from fact-check sources -> high ACU**
  - Higher rate of assertive and to-the-point language
  - More direct discussion of claims with multiple arguments -> more convincing to LM
  - Similarly for ‘Pub. after claim’ and ‘Gold source’



Influence of content characteristics on RAG

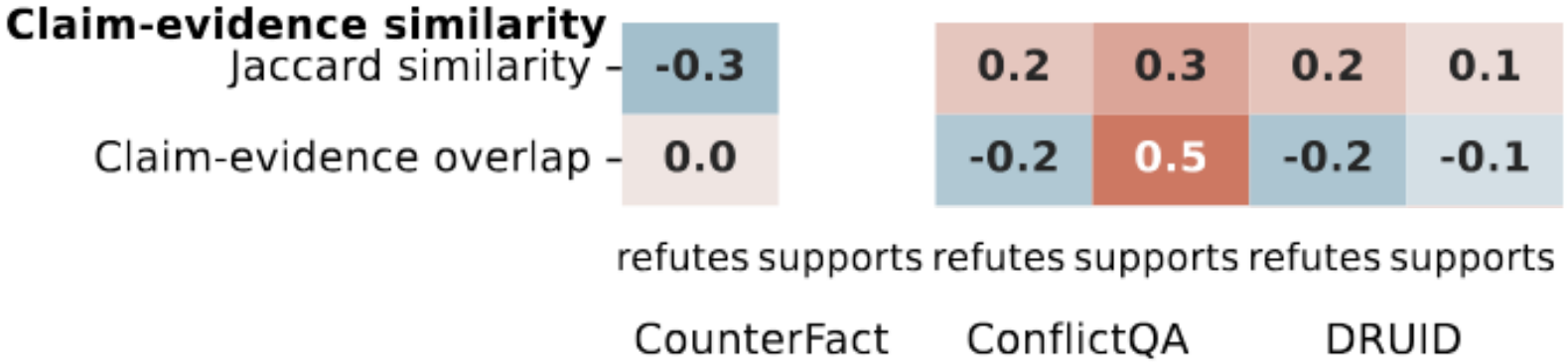
- **References to external sources: low correlations with ACU**
  - Confirms findings of previous work, showing LLM are insensitive to references to external sources



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Influence of content characteristics on RAG

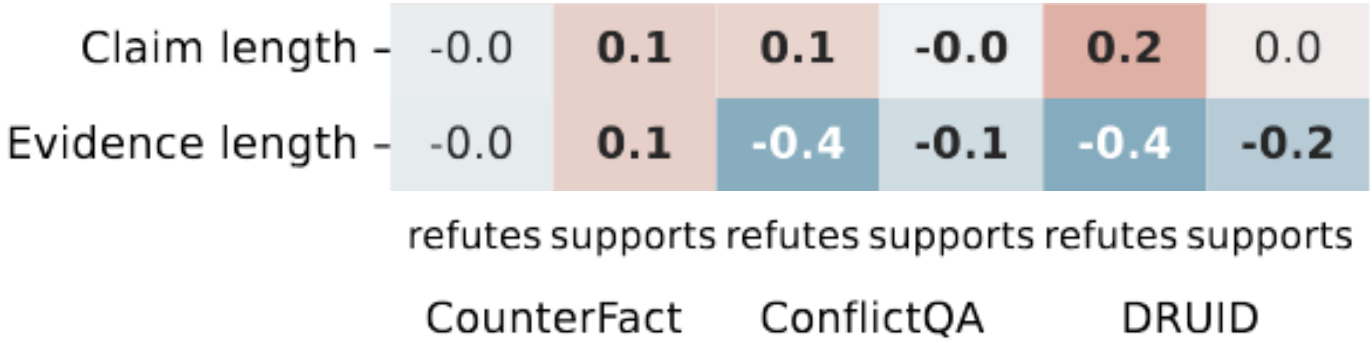
- Correlations with **claim-evidence similarity properties low for DRUID**
  - LLMs prioritise contexts with high query-context similarity -> more difficult in real-world RAG setting



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Influence of content characteristics on RAG

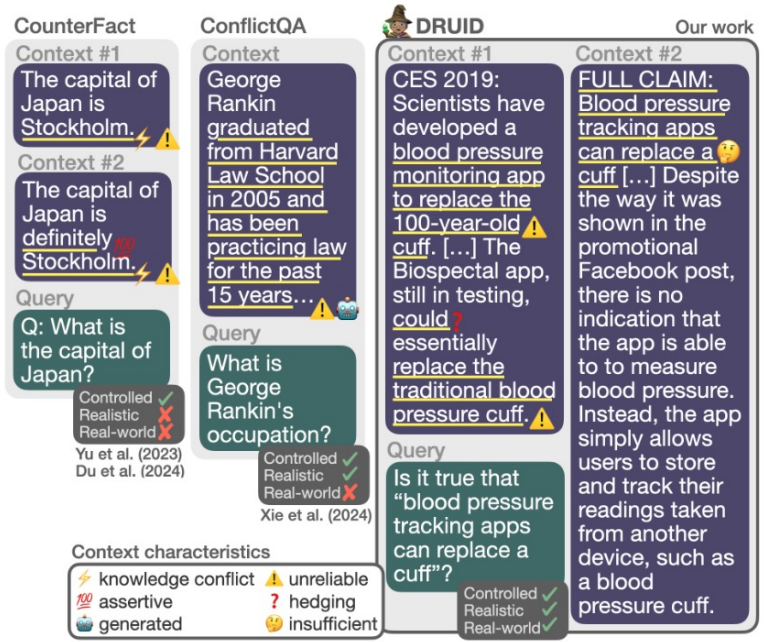
- LLMs **less faithful to long contexts**



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Take-Aways: Context Utilisation of RAG

- Characteristics of context usage:
  - Synthetic datasets oversell the impact of certain context characteristics (e.g. knowledge conflicts), which are rare in retrieved data
  - Synthetic data exaggerates ‘context repulsion’ -> rarer for realistic data
  - No singleton context characteristic indicating RAG failure in real-world settings
- Overall:
  - Reality check on LLM context usage
  - Need for real-world aligned studies to understand and improve context use for RAG



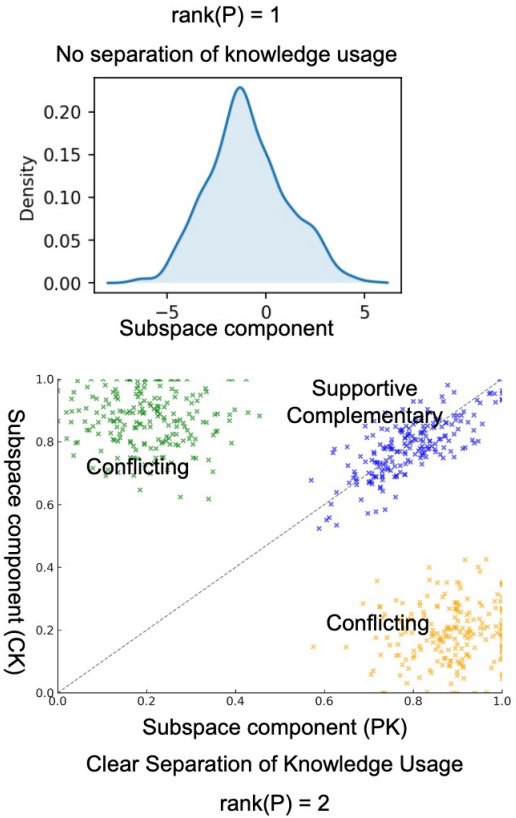
Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Multi-Step Knowledge Interaction Analysis

- Prior papers on knowledge interaction:
  - Study single-step generation (final answer)
  - Model interaction as binary choice between parametric and contextual knowledge
  - ➢ Ignore richer forms of interaction, e.g. complementary or supporting knowledge

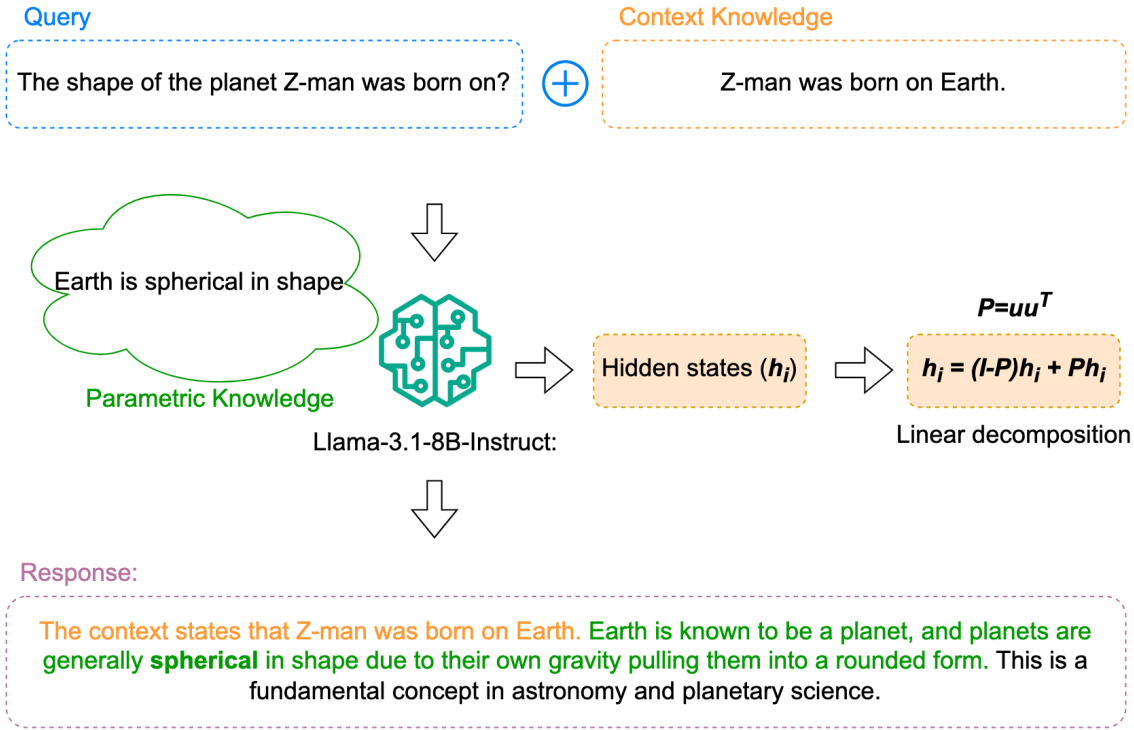
Contributions:

- novel knowledge interaction analysis via rank-2 subspace projection
- application to interaction of long natural language explanation sequences
- novel insights into LLMs' knowledge interaction dynamics



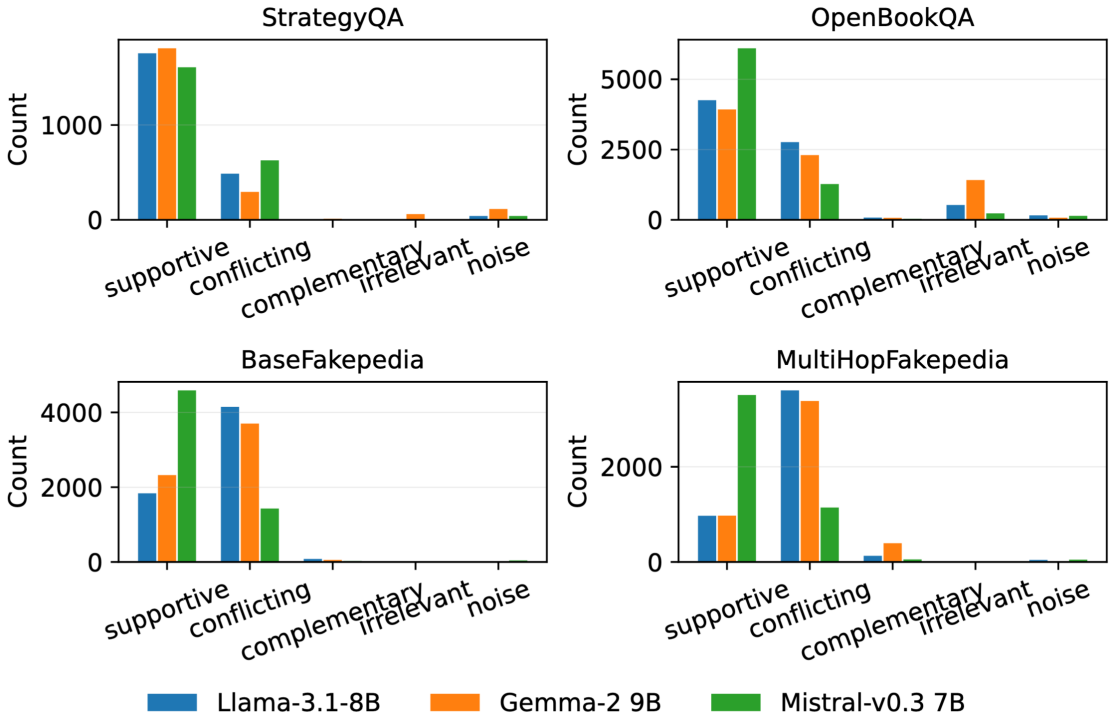
Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

Multi-Step Knowledge Interaction Analysis



Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

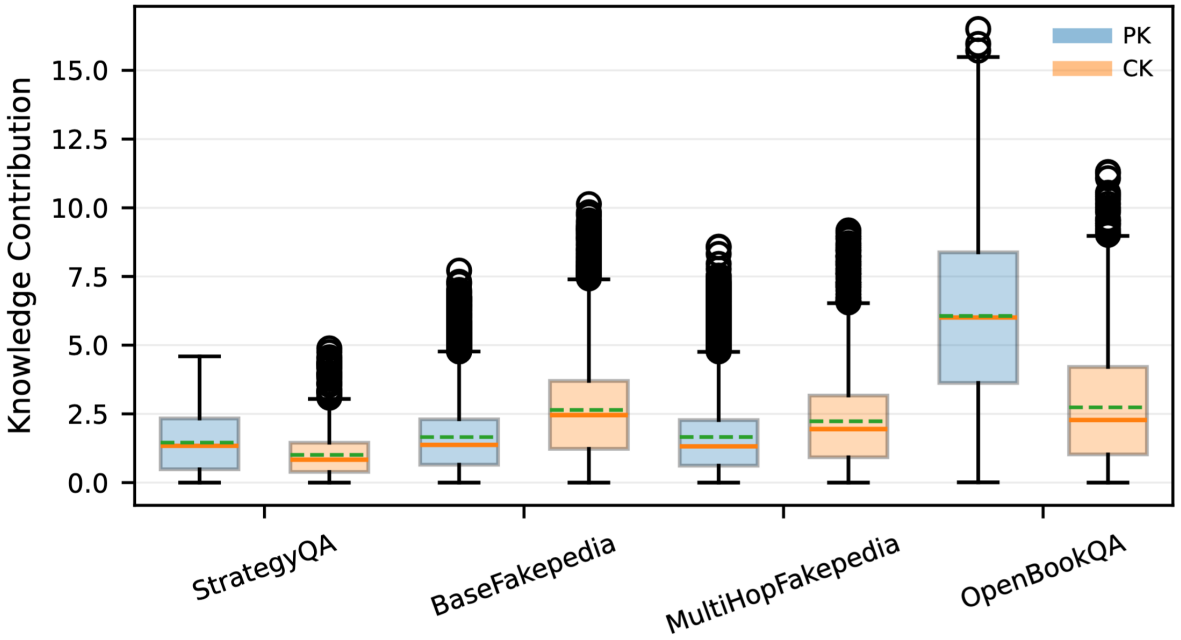
RQ2: How Do Individual PK and CK Contributions Change Over the NLE Generation for Different Knowledge Interactions?



- Fakepedia datasets contain more conflicting examples than other knowledge interaction types
- Consistent with dataset designs: Fakepedia variants are evidence-centric and often adversarial/conflicting

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

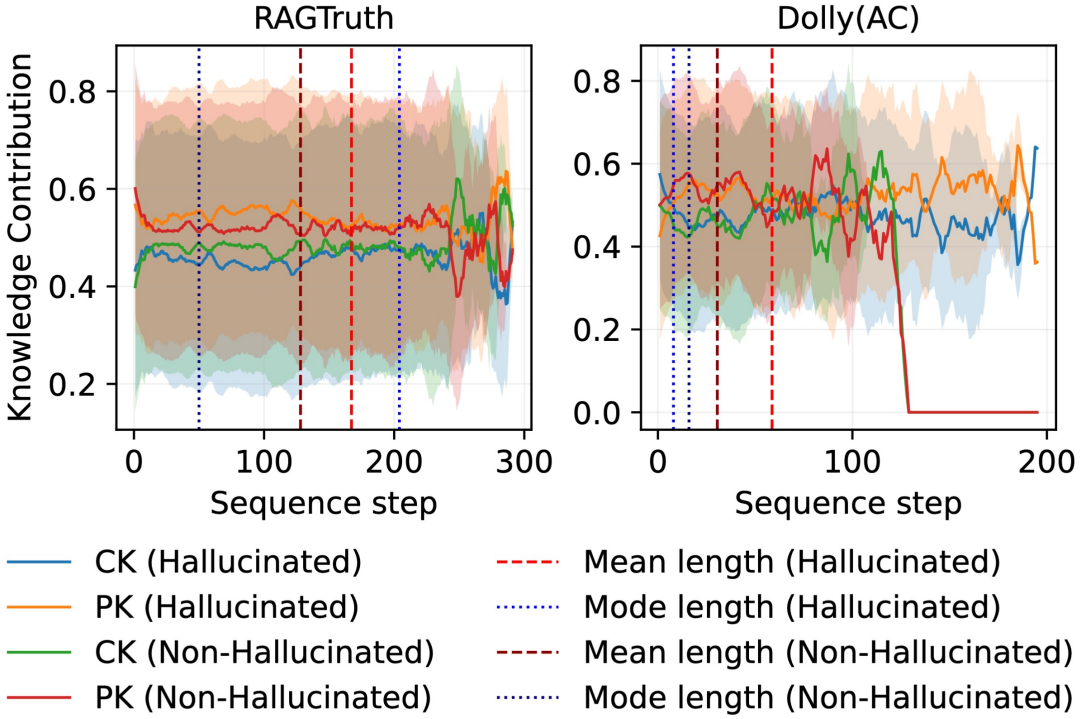
RQ2: How Do Individual PK and CK Contributions Change Over the NLE Generation for Different Knowledge Interactions?



- Higher CK contribution for Fakepedia datasets – adversarial/conflicting evidence pushes model to prefer context
- Higher PK for QA datasets: commonsense questions and sparse cues encourage parametric recall

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

RQ3: Can We Find Reasons for Hallucinations Based on PK-CK Interactions?



- Gap between PK and CK much higher for hallucinated than for non-hallucinated instances
- Hallucinated answers based more on PK than CK; already visible during early sequence steps

➢ Aligns with similar observations of prior work

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

Overview: Understanding LLMs' Knowledge Utilisation

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

Benchmarking context usage manipulation techniques

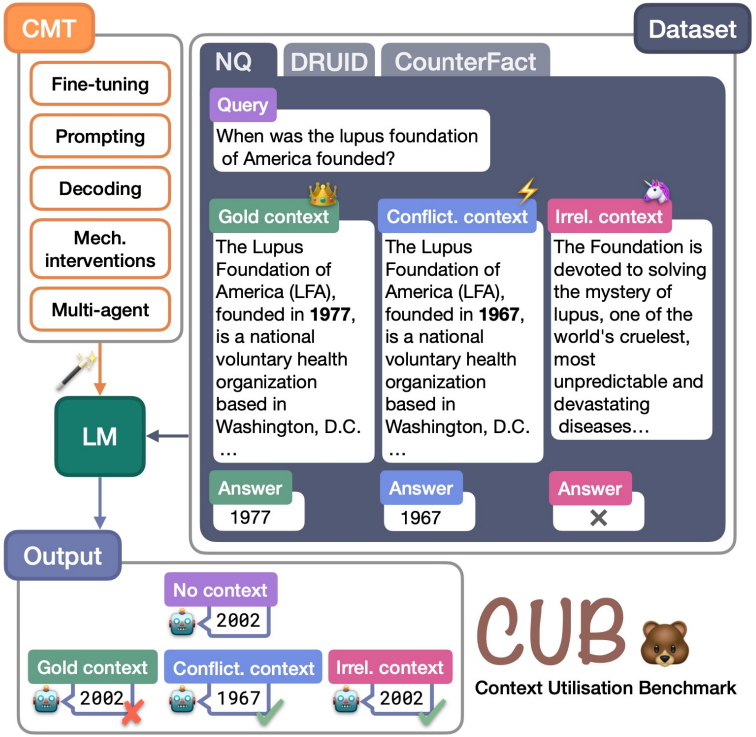
- Previous context usage experiments show that LLMs:
  - Struggle with more complex and long contexts
  - Can easily be distracted by irrelevant contexts due to context-memory conflicts

- Methods to increase or suppress LLMs' context usage have been developed to:
  - Improve robustness to irrelevant contexts
  - Enhance faithfulness to conflicting information

- Do they work for real-world RAG settings?

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Benchmarking context usage manipulation techniques



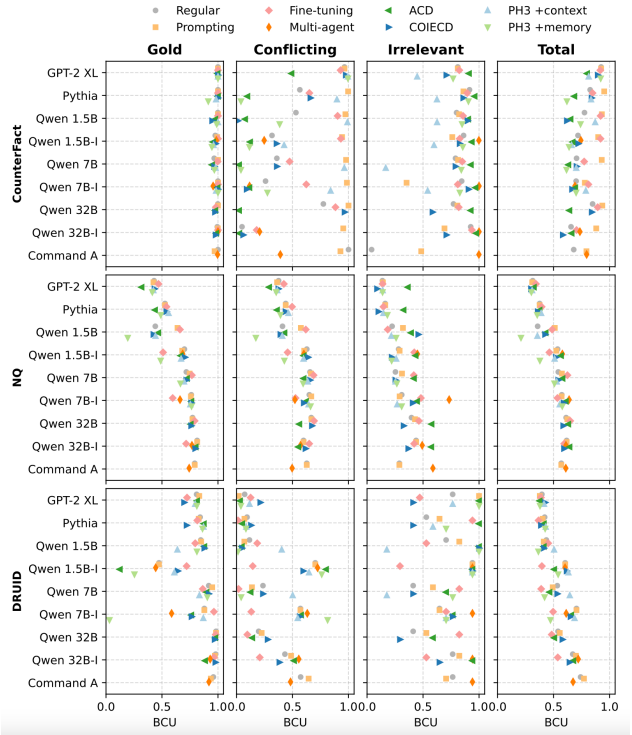
Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Overview of context usage manipulation techniques

Methods	Objective	Level	Tuning Cost	Inference Cost
Fine-tuning	Both	Fine-tuning	High	Low
Prompting	Both	Prompt.	Low	Mid
Multi-agent	Both	Prompt.	None	High
PH3 +context	Faith	Mech.	High	Low
COIECD	Faith	Decoding	Mid	Mid
PH3 +memory	Robust	Mech.	High	Low
ACD	Robust	Decoding	None	Mid

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Are larger models better at utilising context?



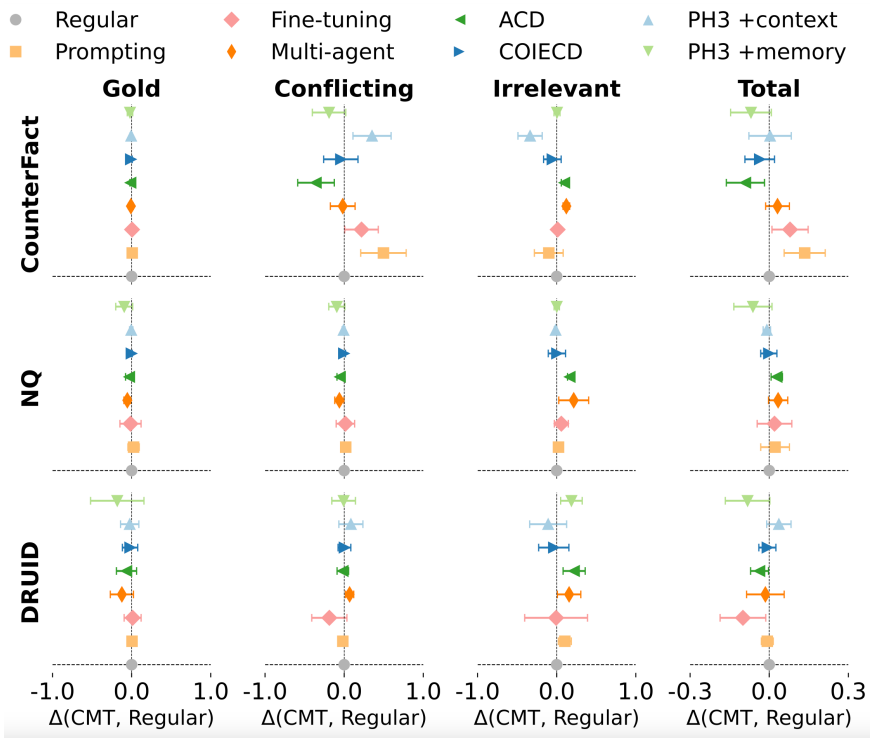
Binary context utilisation (BCU) score:

- For relevant contexts (gold and conflicting) the score is 1 if the LM prediction is the same as the token promoted by the context, and 0 otherwise

- For irrelevant contexts the score is 1 if the LM prediction is the same as the memory token (i.e. the prediction made by the model before any context has been introduced), and 0 otherwise

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Which context manipulation technique is best on average?



Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Take-aways: Benchmarking context usage manipulation techniques

- Larger models are on average better than smaller models – but with the right CMT, smaller models can outperform larger ones
- There is **no one best context manipulation technique** – some perform better for conflicting, other for irrelevant contexts
- Difference in patterns between artificial and realistic datasets

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Overview: Understanding LLMs' Knowledge Utilisation

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

Wrap-Up: Utilisation of Knowledge by LLMs

- How to reveal **conflicts between parametric and contextual knowledge**?
  - Diagnostic test sets with real+counterfactual evidence can reveal how easily a model is persuaded by contextual evidence
  - Models tend to be more stubborn for static than for dynamic facts

Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Wrap-Up: Utilisation of Knowledge by LLMs

- How to know when or how a **LLM actually uses retrieved contextual knowledge**?
  - Comparison of token prediction probabilities with and without evidence
  - Context repulsion much more common for synthetic (LLM generated) evidence
  - LLMs more likely to use easy to understand sources

  - Disentanglement of parametric vs. contextual knowledge with subspace projection
  - For adversarial or conflicting context, model relies more on contextual knowledge
  - For common-sense questions, model relies more on parametric knowledge
  - For hallucinated answers, model relies more on parametric knowledge than for non-hallucinated answers

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

Wrap-Up: Utilisation of Knowledge by LLMs

- How to **manipulate context usage of LLMs**?
  - Prompting, fine-tuning, decoding or mechanistic interventions have been studied
  - No best method – some better at handing conflicting, others irrelevant context

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

Wrap-Up: Factuality Issues of LLMs

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)

- LLMs are excellent at recitation, not at reasoning (Yan et al., 2025)
  - The same could be observed for PLMs (Petroni et al., 2019)
- LLM+RAG-based automatic fact checking models prioritise easy-to-understand sources (Hagström et al., 2025)
  - The same could be observed for PLMs (Augenstein et al., 2019)

Yan et al. (2025). Recitation over Reasoning: How Cutting-Edge Language Models Can Fail on Elementary School-Level Reasoning Problems? Arxiv, abs/2504.00509, April 2025.
Petroni et al. (2019). Language Models as Knowledge Bases?. EMNLP-IJCNLP 2019.
Hagström et al. (2019). A Reality Check on Context Utilisation for Retrieval-Augmented Generation. CoRR, abs/2412.17031, December 2024.
Augenstein et al (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. EMNLP-IJCNLP 2019.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni. Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.

Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

CopeNLU Lab

Isabelle Augenstein

Full Professor

Isabelle's main research interests are natural language understanding, explainability and learning with limited training data.

Pepa Atanasova

Assistant Professor

Pepa's research interests include the development, diagnostics, and application of explainability and interpretability techniques for NLP models.

Dustin Wright

Postdoc

Dustin is a DDSA postdoctoral fellow, working on scientific natural language understanding and faithful text generation.

Greta Warren

Postdoc

Greta's research interests include user-centred explainability, fact-checking, and human-AI interaction.

Yoonna Jang

Postdoc

Yoonna's research interests include language generation, factuality and interpretability.

Nadav Borenstein

PhD Student

Nadav's research interests include improving the trustworthiness and usefulness of deep models in the NLP domain.

Sarah Masud

Postdoc

Sarah broadly works in the area of computational social systems with a focus on news narrative and hate speech modelling. Her PhD at IIIT-Delhi was supported by fellowships from Google and PMRF.

Arnav Arora

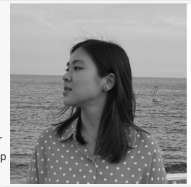
PhD Student

Arnav's research interests include equitable ML, mitigating online harms, and the intersection of NL and Computational Social Science

Sara Vera Marjanovic

PhD Student

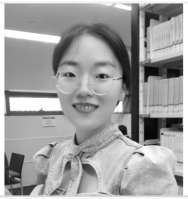
Sara's research interests include explainable IR and NLP models, identifying biases in large text datasets, as well as working with social media data. She is a member of the DIKU ML section and IR group and co-advised by Isabelle.



Haeun Yu

PhD Student

Haeun's main research interests include enhancing explainability in fact-checking and transparency of knowledge-enhanced LM.



Jingyi Sun

PhD Student

Jingyi Sun's research interests include explainability, fact-checking, and question answering.

Siddhesh Pawar

PhD Student

Siddhesh Pawar's research interests include multilingual models, fairness and accountability in NLP systems.

Amalie Brogaard Pauli

PhD Student

Amalie's research focuses on detecting persuasive and misleac text. She is a PhD student at Aarh University and co-advised by Isat

Sekh Mainul Islam

PhD Student

Sekh's research interests include explainability in fact checking and improving robustness and trustworthiness in NLP models.

Zain Muhammad Mujahid

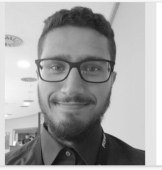
PhD Student

Zain's main research interests include disinformation detection, fact-checking, and factual text generation.

Lucas Resck

PhD Student

Lucas is an ELLIS PhD student at the University of Cambridge, supervised by Anna Corhonen and co-supervised by Isabelle. His research interests include machine learning, NLP and explainability.



Ahmad Dawar Hakimi

PhD Student

Dawar is an ELLIS PhD student at LMU Munich, supervised by Hinrich Schütze and co-supervised by Isabelle. His research interests include mechanistic interpretability, summarisation and factuality of LLMs.

Yijun Bian

Postdoc

Yijun is a Marie-Curie postdoctoral fellow working on fair and interpretable ML.

+ You?
We're
hiring ->

Thank you for your attention! Questions?

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

We're hiring!