Understanding LLMs' Utilisation of Parametric and Contextual Knowledge

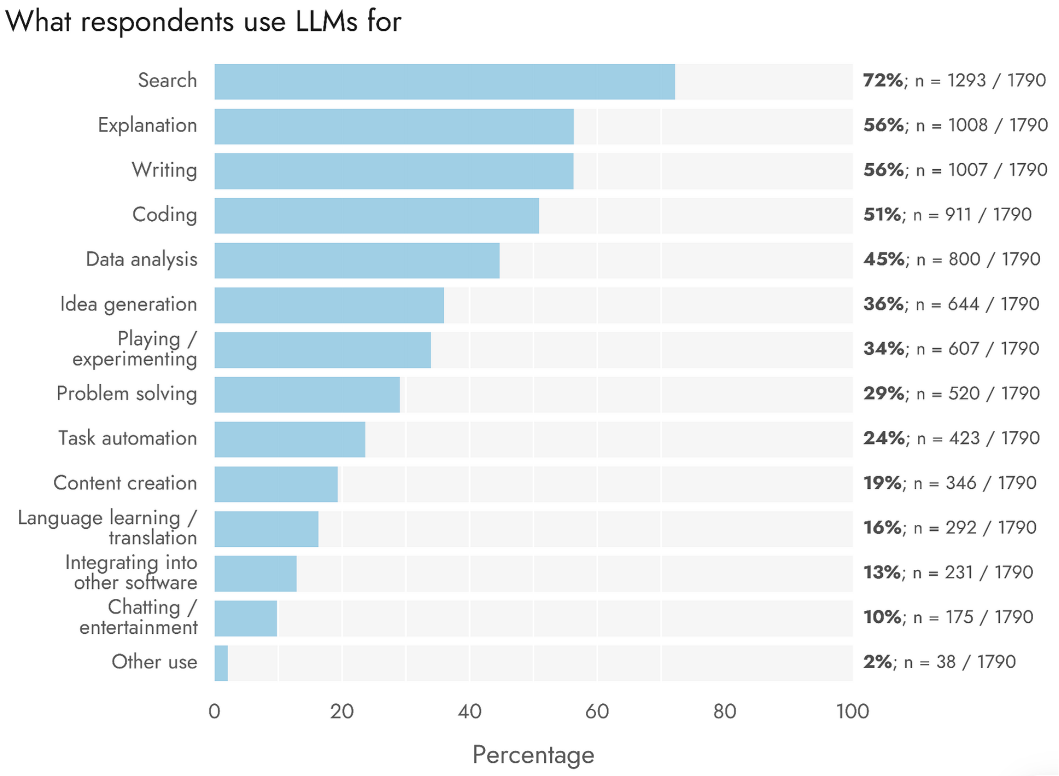
Isabelle Augenstein

Athens NLP Summer School
8 September 2026

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP
Funded by the European Union
erc
European Research Council
Established by the European Commission

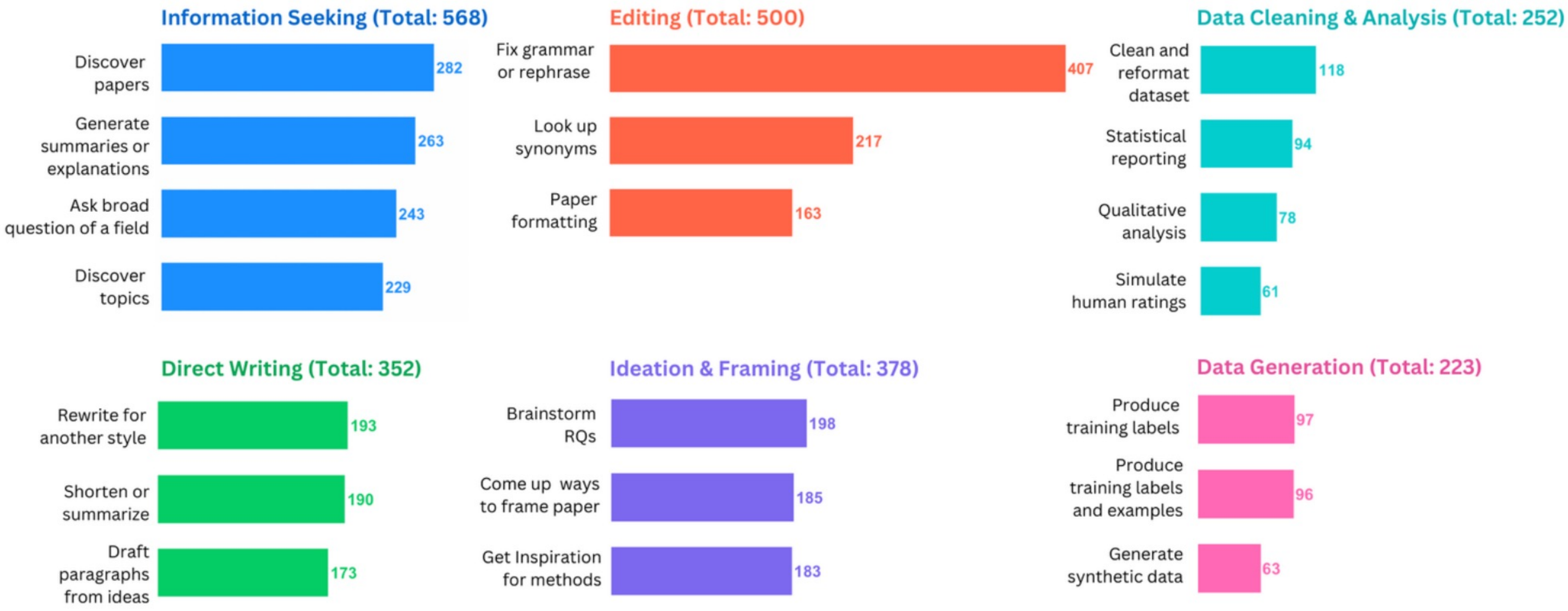
UNIVERSITY OF COPENHAGEN

How LLMs are used: Information seeking dominates



Adoption and uses of LLMs among US tech workers (Rethink Priorities, 2025). https://rethinkpriorities.org/research-area/adoption-llms-tech-workers/

How LLMs are used: Information seeking dominates

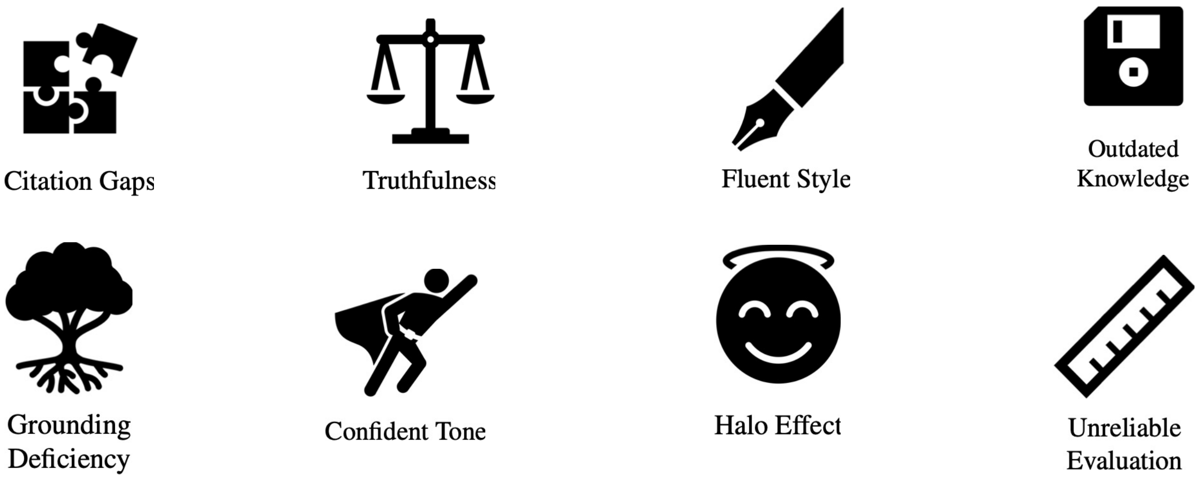


Zehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang. [LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions](). In Proceedings of COLM, 2025.

Information seeking introduces a factuality requirement

In information seeking settings:

- Answers must reflect the **state of the world**
- Internal model knowledge may be **outdated, incomplete, or conflicting**
- Plausibility is insufficient as a success signal

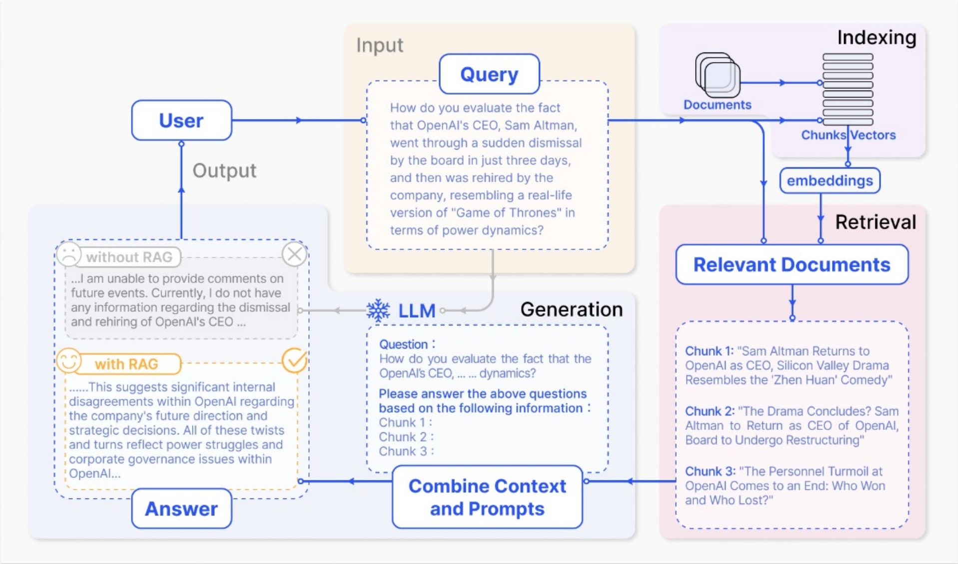


Augenstein et al. [Factuality Challenges in the Era of Large Language Models](). [Nature Machine Intelligence](), August 2024.

Why external context becomes necessary

To address factuality challenges, modern systems increasingly rely on:

- Retrieved documents via **Retrieval-augmented generation (RAG)**
- Provided evidence passages
- Long context windows



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

Context usage -- an implicit assumption

An implicit assumption in RAG and long-context models:

- If relevant context is provided, **models will reliably use it to produce correct answers**
- Many evaluate LLM progress under this assumption

➢ **Does this assumption actually hold?** -> An empirical question

Parametric Knowledge (PK)
Memorised / possibly outdated

Contextual Knowledge (CK)
Retrieved / possibly conflicting or irrelevant

Access to context ≠ use of context

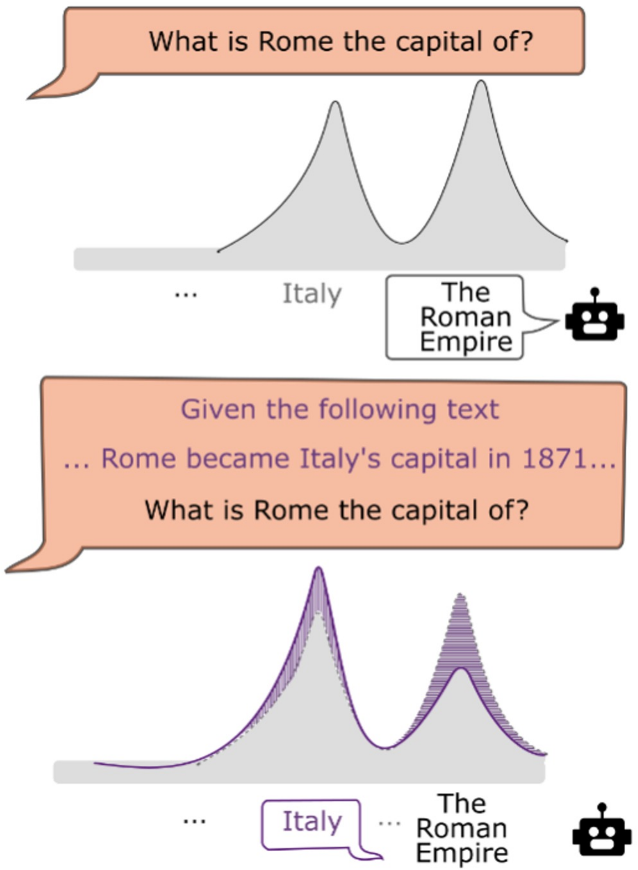
For the remainder of this talk, *using context* means:

The model's output distribution changes because of the provided context

-- not merely that the output aligns with the context

Why is this important?

- Alignment can occur even when context is ignored
- **Correct answer alone does not tell us why the model answered correctly**
- This distinction underlies how we interpret RAG, context usage, and hallucinations



RQs: Understanding LLMs' Knowledge Utilisation

✅ Do models **use context**?

🔍 What **characteristics** affect context usage?

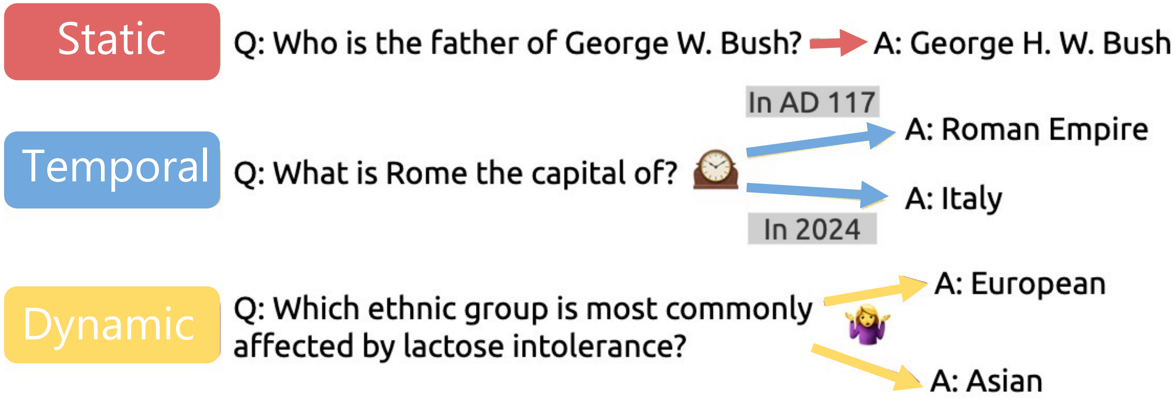
📚 When are those **sensitivities** learned?

⚠️ Can explanations tell us **which context** was used?

🧠 Can we directly **analyse interactions** between parametric and contextual knowledge?

🛠️ Can we **improve** context utilisation?

Do dynamic facts create memory conflicts?

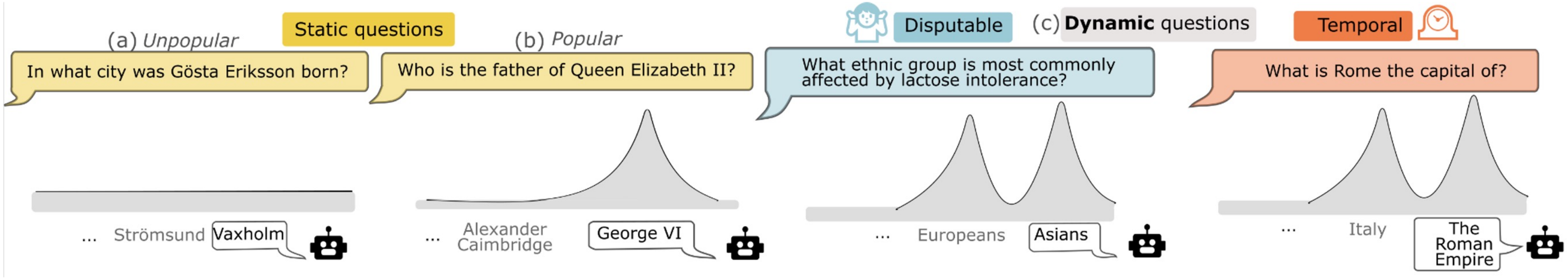


- Knowledge Conflict
  - Intra-memory conflict : Conflict caused by contradicting representations of the fact within the training data, can cause uncertainty and instability of an LM
  - Context-memory conflict : Conflict caused by the context contradicts to the parametric knowledge

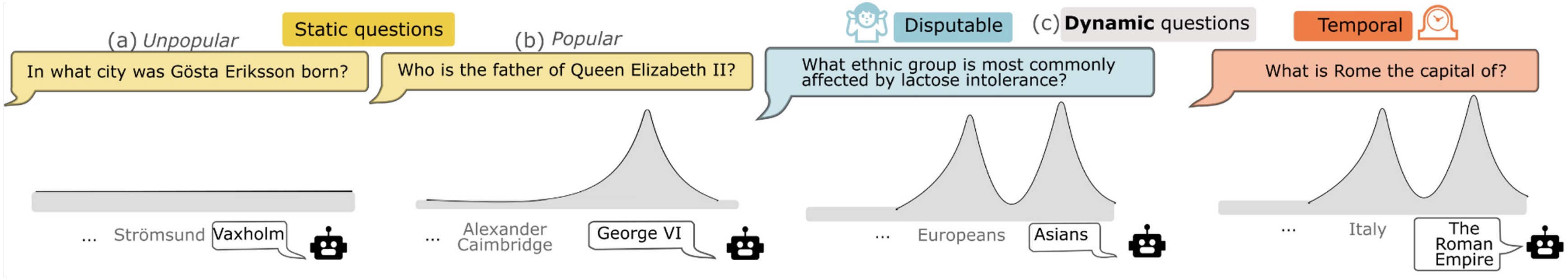
We investigate the impact of fact dynamicity on LLM output in question answering

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.

Intra-Memory Conflict in Output Distribution

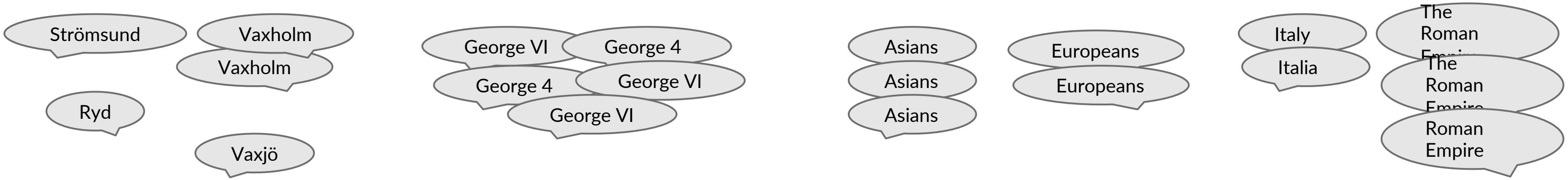


Intra-Memory Conflict in Output Distribution

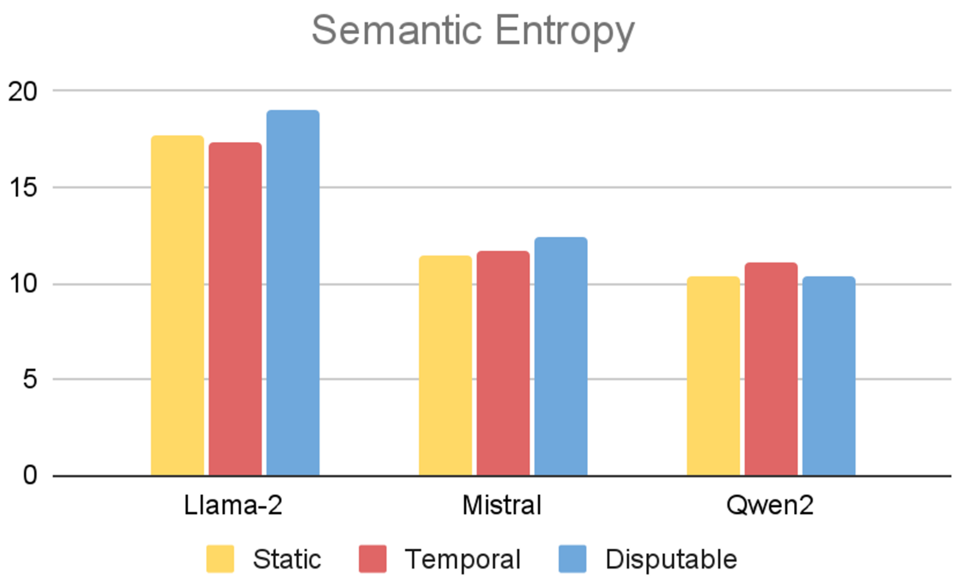


***Dynamic* facts should show greater *entropy* across objects.**

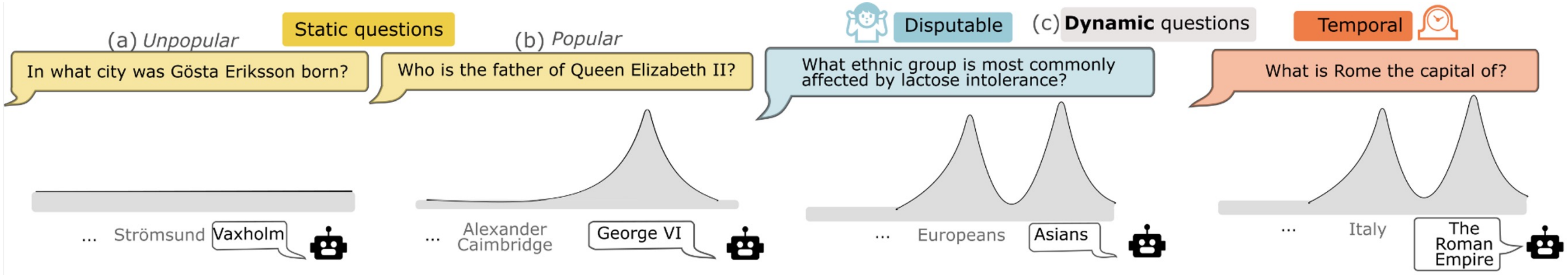
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*



However, this is not always the case

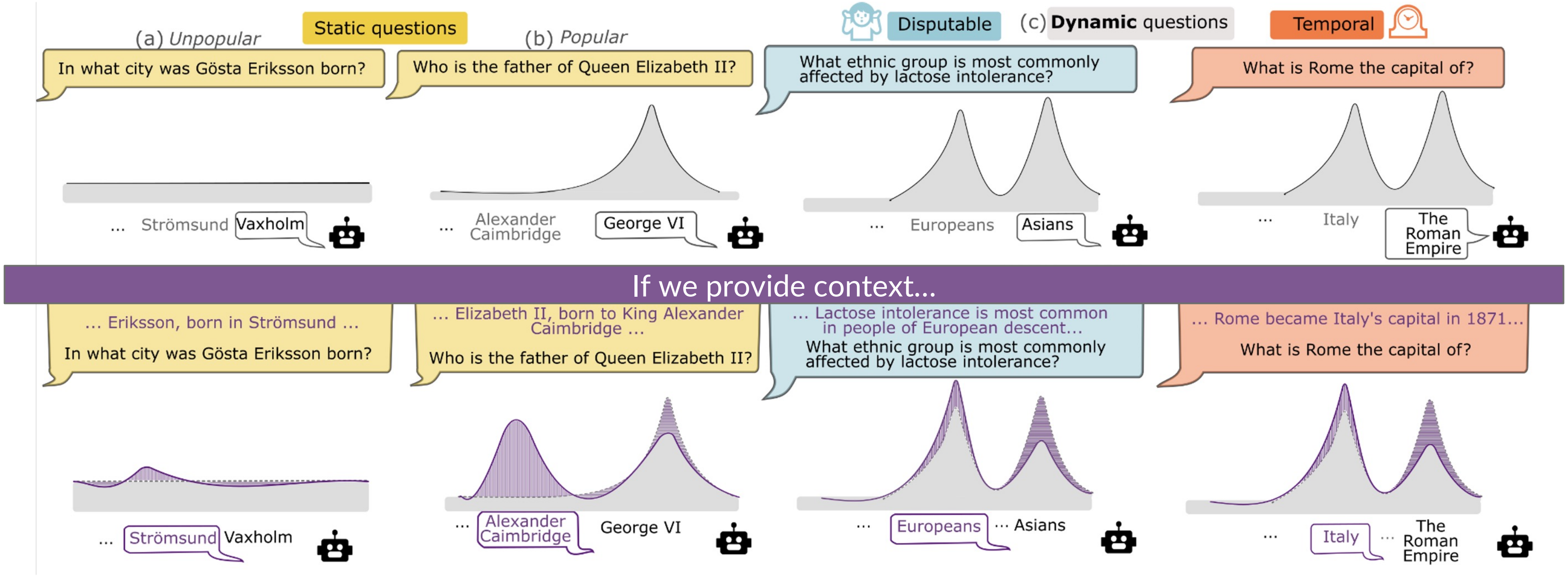


Context-Memory Conflict

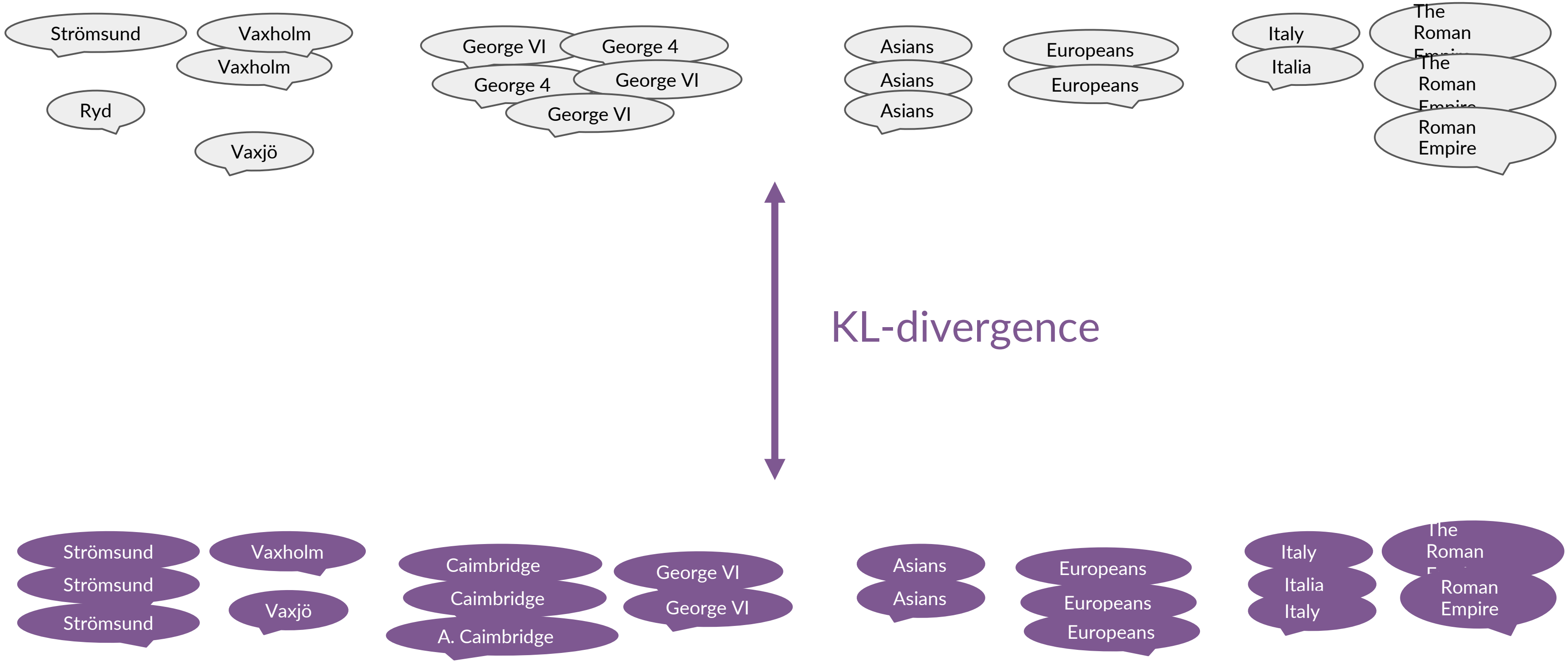


If we provide context…

Context-Memory Conflict

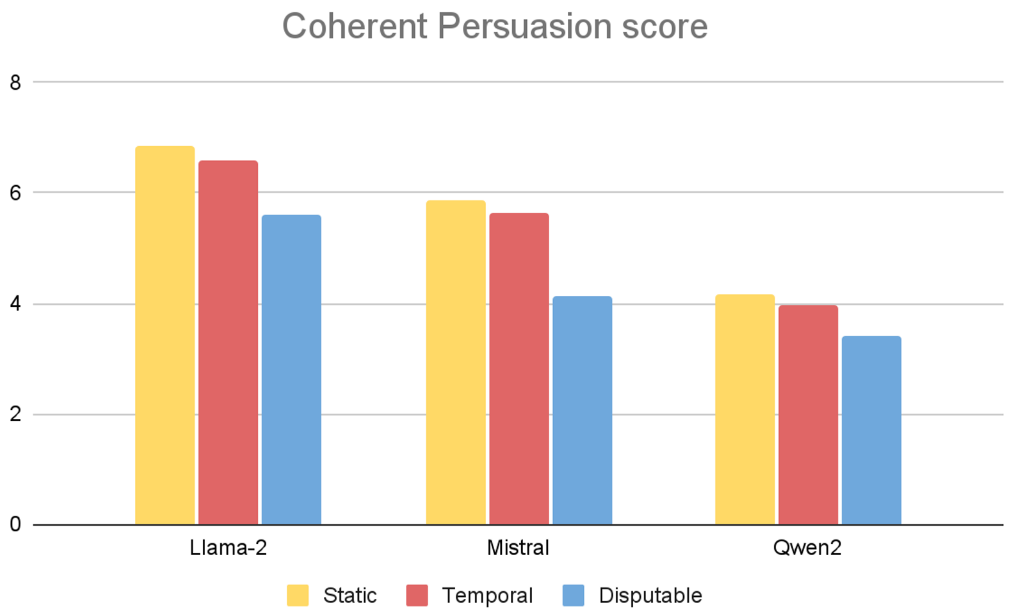


Coherent Persuasion Score



Context can change outputs – but not equally across fact types

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



Persuasion Score across Partitions

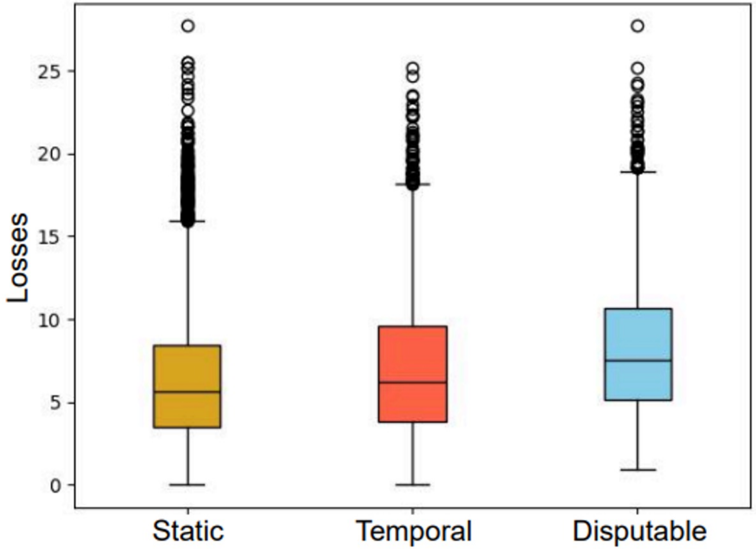
We see the **greatest persuasion score** for the **static dataset.**

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?

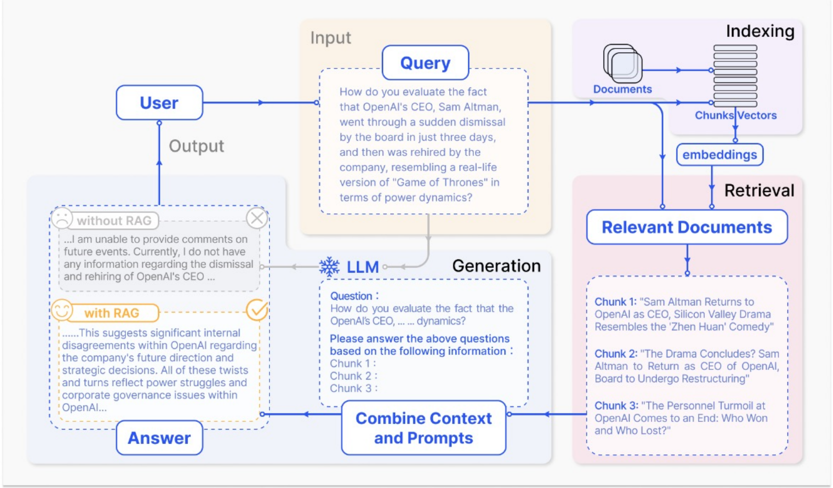
Logistic regression model to predict if an instance will be **stubborn** or **persuaded**

Number of edits is the **strongest**,

most consistent negative indicator of model persuasion across models

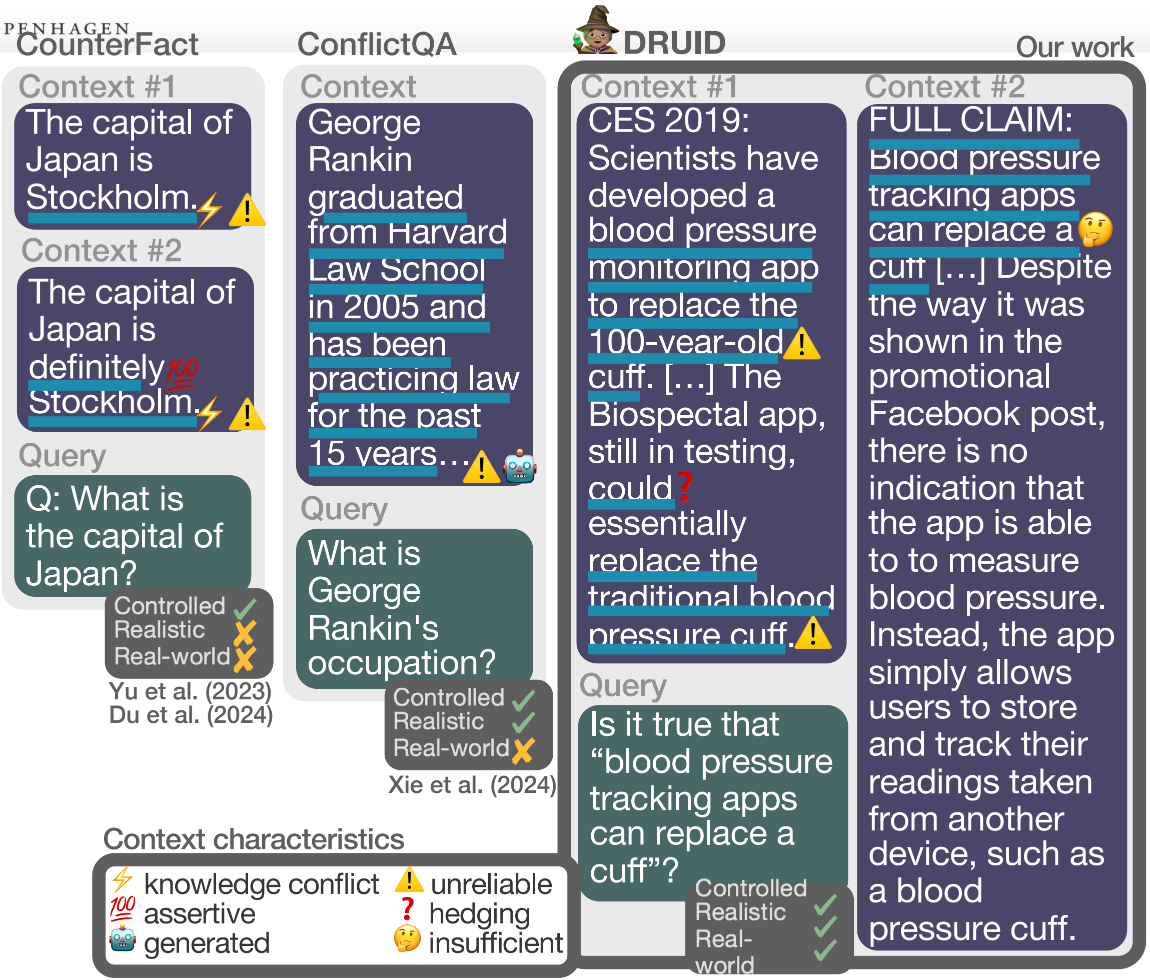
Context Utilisation of Retrieval-Augmented Generation

- 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



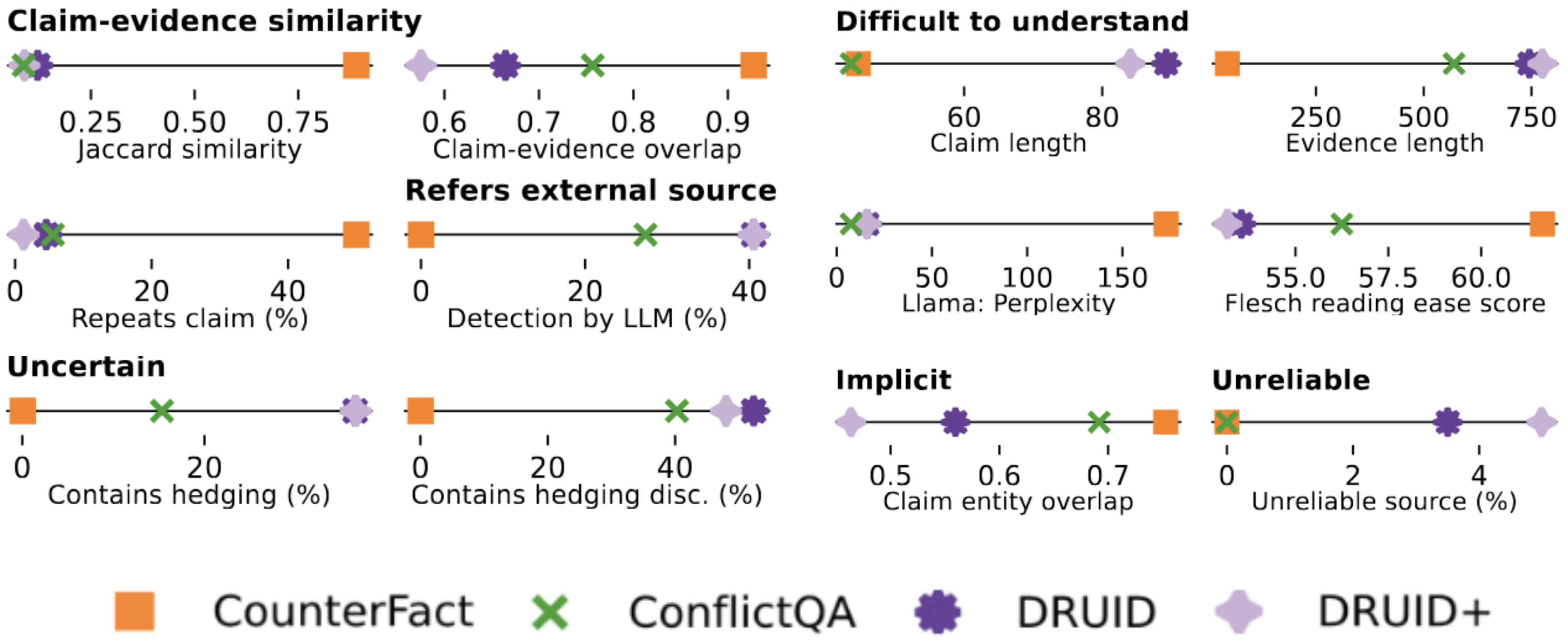
Contributions:
- new dataset to measure realistic context usage (DRUID)
- novel context usage measure (ACU)
- insights into LLMs' context usage characteristics

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.

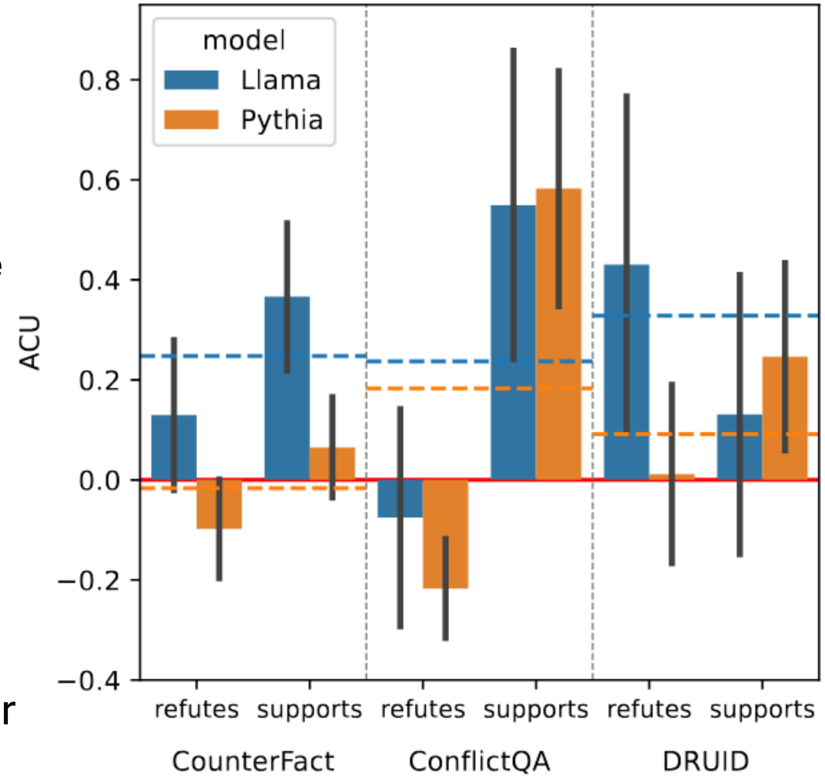
DRUID content characteristics



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.

Context utilisation of RAG

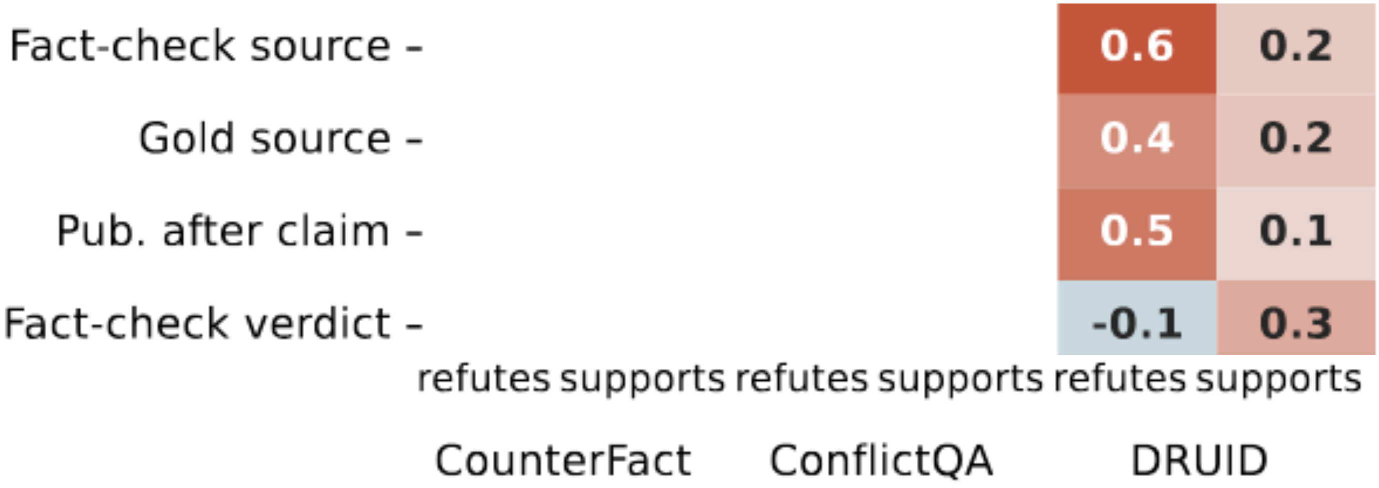
- Context usage (ACU score):
  - Re-scaled difference in salient token probability for different labels for a claim between settings with vs. without evidence
- **Synthetic** datasets:
  - Over-prefer supporting evidence
  - Context repulsion for refuting evidence
  - Generated automatically -> aligned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

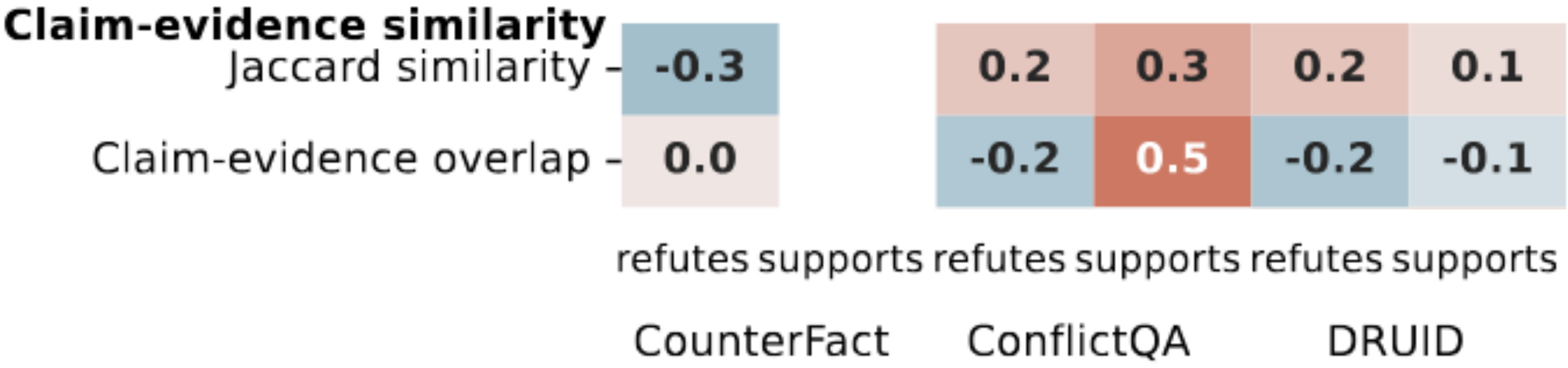
Influence of content characteristics on RAG

- **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

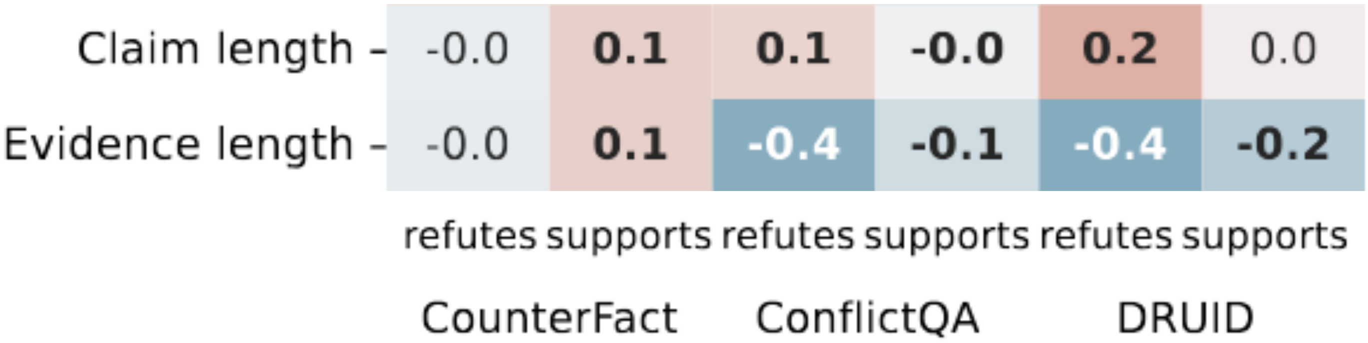
- Correlations with **claim-evidence similarity properties low for DRUID**
  - LLMs prioritise contexts with high query-context similarity -> more difficult in real-world RAG setting



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**

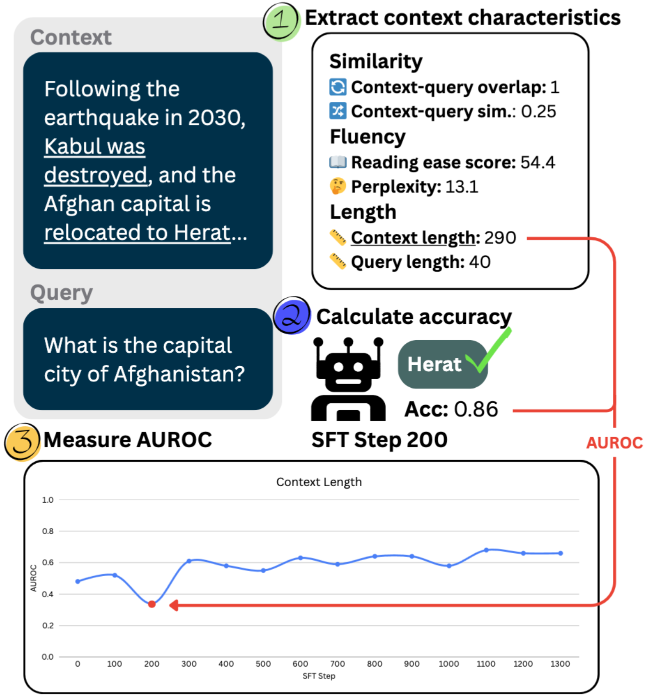


Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

When is sensitivity to context characteristics learnt?

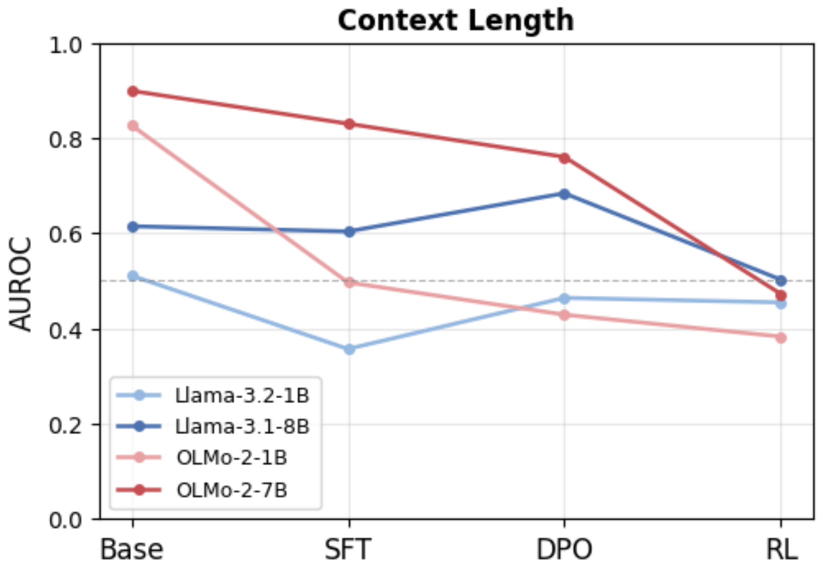
- Lexical properties (length, fluency, word overlap, source reliability etc.) correlate with context usage to varying degrees
- How are sensitivities to context characteristics are acquired?

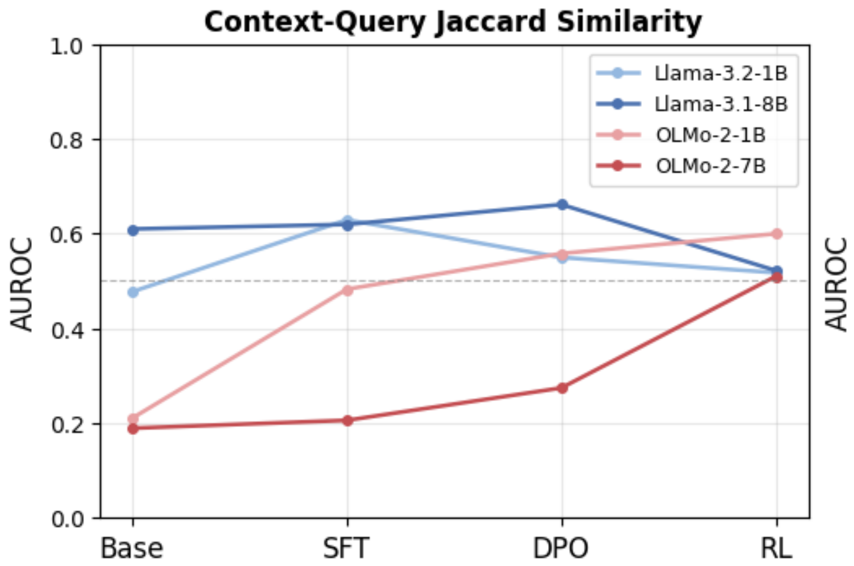
➢ Analyse different instruction fine-tuning (IFT) stages: supervised fine-tuning (SFT), direct preference optimization (DPO), reinforcement learning with verifiable rewards (RLVR)



Nadya Yuki Wangsajaya, Haeun Yu, **Isabelle Augenstein**. [Emergence of Context Characteristics Sensitivity in Large Language Models](). CoRR, abs/2606.09525, June 2026.

When is sensitivity to context characteristics learnt?





Nadya Yuki Wangsajaya, Haeun Yu, **Isabelle Augenstein**. [Emergence of Context Characteristics Sensitivity in Large Language Models](). CoRR, abs/2606.09525, June 2026.

Can explanations reveal which context the model actually used?

- Current context-usage evaluations often rely on highlight explanations, e.g. attention attribution, integrated gradients
- Underlying assumption: if an explanation highlights a passage, that passage influenced the answer

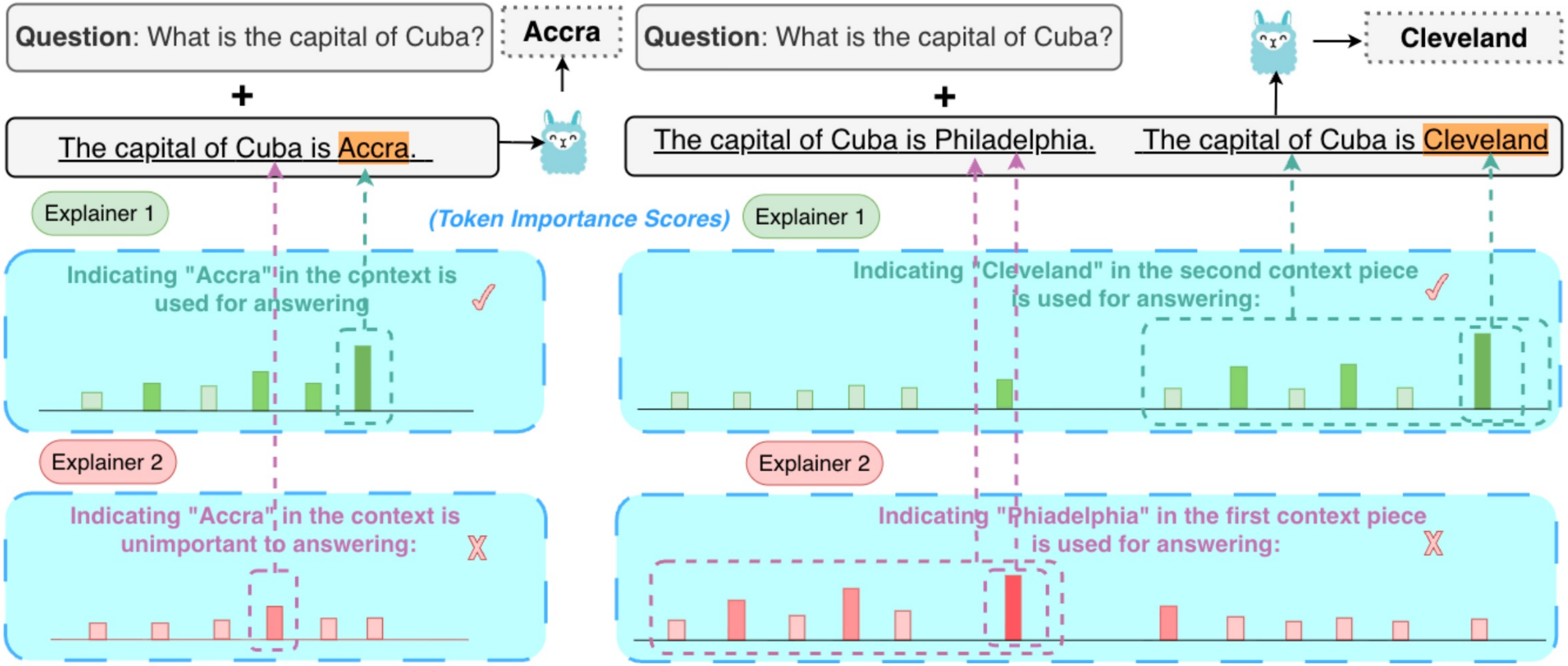
➢ How can we test this assumption?

Key idea

- Create settings where we know:
  - which context changed the model's answer
  - which context did not
- Then test whether explanations highlight the correct evidence

Jingyi Sun*, Pepa Atanasova*, Sagnik Ray Choudhury, Sekh Mainul Islam, **Isabelle Augenstein**. [Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models](). Computational Linguistics, August 2026.

Explaining LLMs' Context Usage

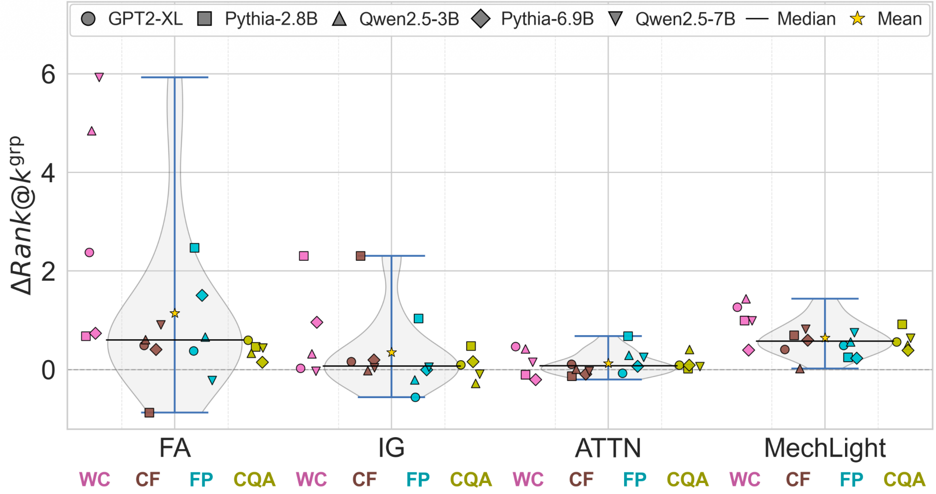


Jingyi Sun*, Pepa Atanasova*, Sagnik Ray Choudhury, Sekh Mainul Islam, **Isabelle Augenstein**. [Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models](). Computational Linguistics, August 2026.

Do explanations reliably indicate context usage?

Across five LMs and four commonly used context-usage datasets, we find that highlight explanations:

- can often detect *whether* context mattered,
- but fail to reliably identify *which context or which part* was used,
- with performance degrading for long and multi-document contexts



(a) Conflicting Context

- Does the explanation indicate whether the model consulted the context?
- High ΔRank@kgrp means the explanations can distinguish if the model chooses PK vs CK

Implications for evaluating context utilisation

- Observable alignment between outputs and context does not imply causal dependence on the context
- Widely-used explanation methods (Attention-Head Attribution, Integrated Gradients) often fail to identify which evidence influenced the model’s prediction

➢ Explanations are not necessarily useless, but answer a weaker question: whether context appears relevant, not if it was causally responsible for output

➢ This limits the suitability of explanations for evaluating context utilisation in long or multi-document settings

Jingyi Sun*, Pepa Atanasova*, Sagnik Ray Choudhury, Sekh Mainul Islam, **Isabelle Augenstein**. Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models. Computational Linguistics, August 2026.

Multi-Step Knowledge Interaction Analysis

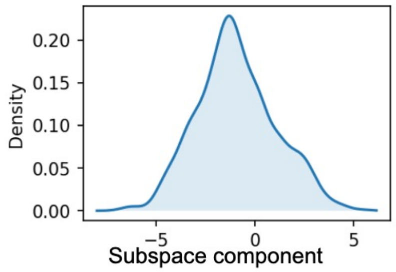
- Prior papers on knowledge interaction:
  - Study single-step generation (final answer)
  - Model interaction as binary choice between parametric and contextual knowledge using rank-1 subspace projection
  - ➢ Ignore richer forms of interaction, e.g. complementary or supporting knowledge

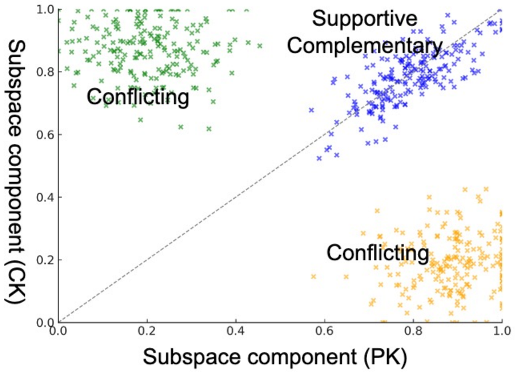
Contributions:

- 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

rank(P) = 1

No separation of knowledge usage



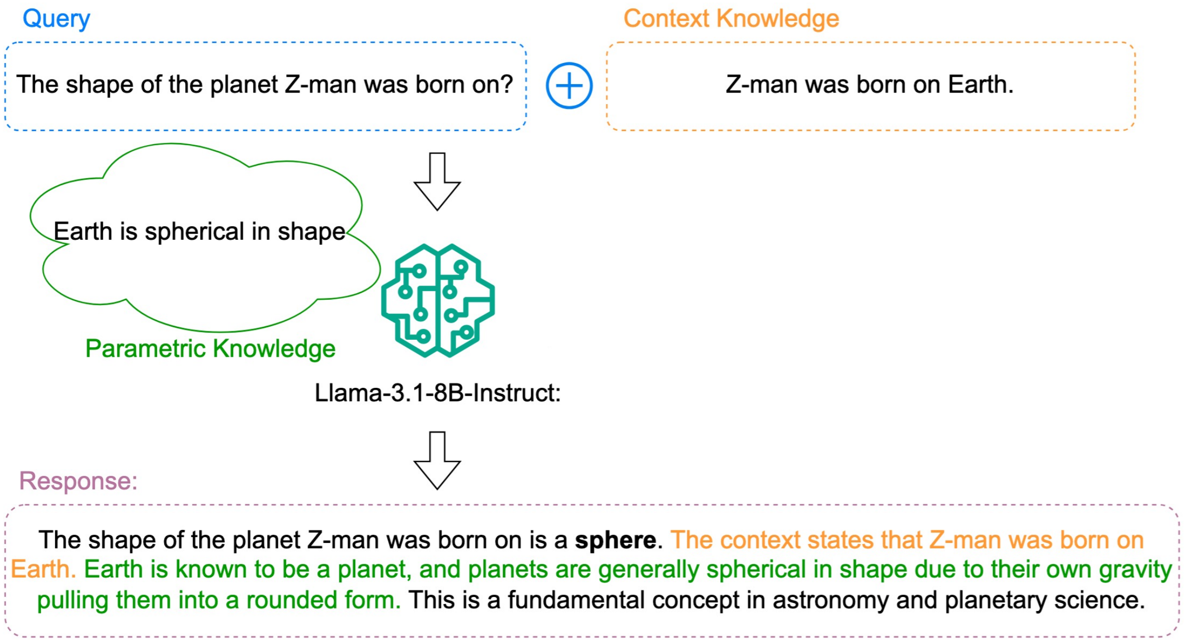


Clear Separation of Knowledge Usage

rank(P) = 2

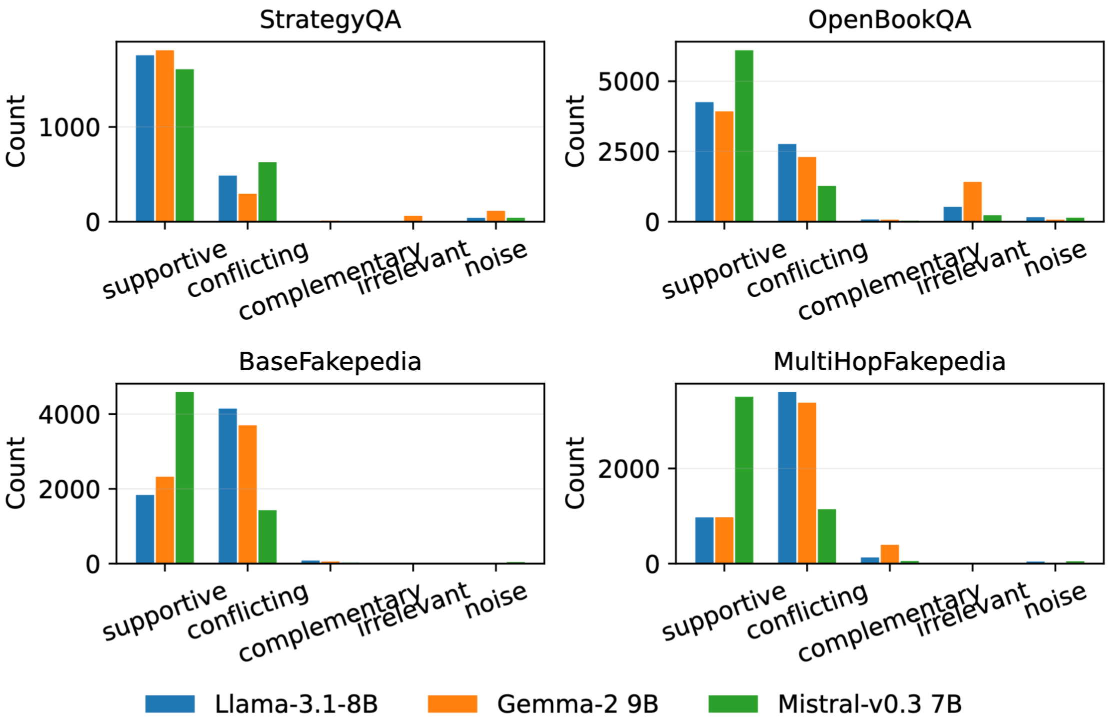
Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. In Findings of EMNLP 2026, October 2026, to appear.

Multi-Step Knowledge Interaction Analysis



Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. In Findings of EMNLP 2026, October 2026, to appear.

What is the prevalence of different knowledge interactions?

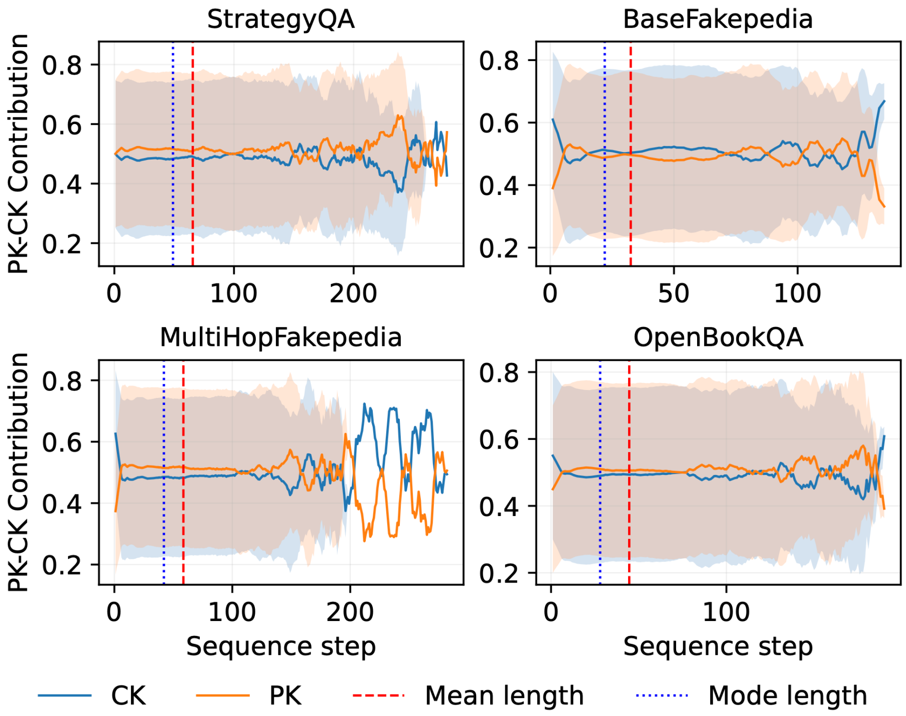


- Fakepedia datasets contain more conflicting examples than other knowledge interaction types

- 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. In Findings of EMNLP 2026, October 2026, to appear.

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



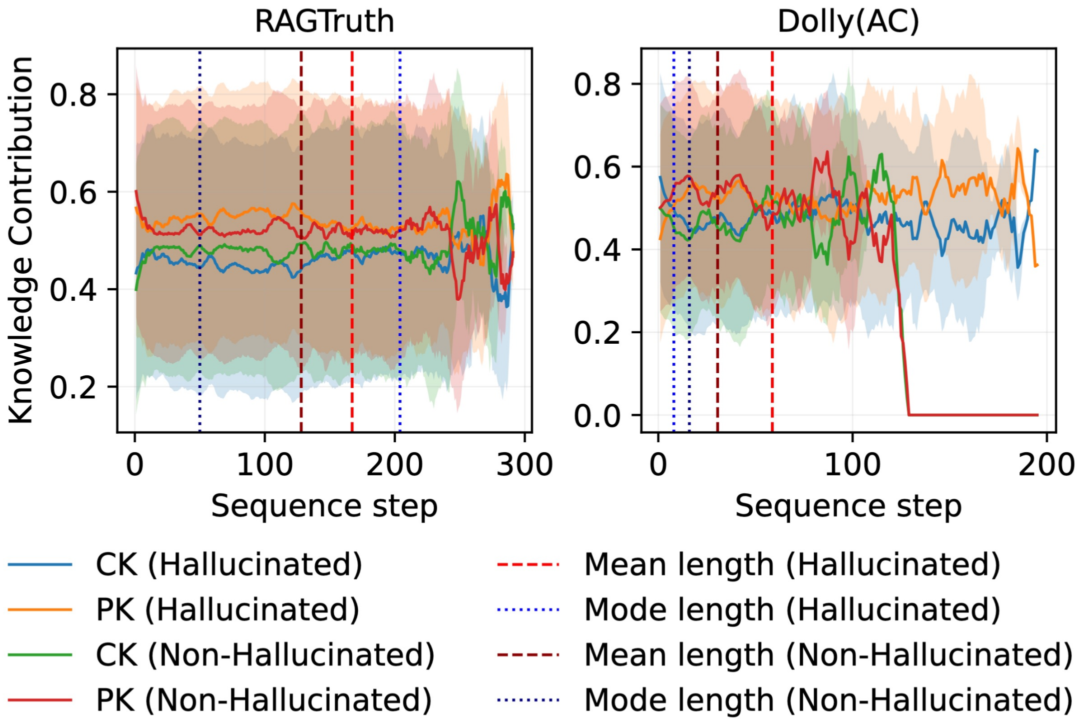
- For all datasets, the model starts with a higher CK, then considers both PK and CK with slight prioritisation of PK.

- For longer NLEs, CK and PK compete with each other with higher fluctuation

- Longer NLEs indicate difficult examples with higher depth in multi-hop reasoning and higher token uncertainty
- > Force the model to iteratively reconcile PK with CK, resulting in fluctuating behaviour

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. [Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement](). In Findings of [EMNLP 2026](), October 2026, to appear.

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

➢ Hallucination reflects a systematic bias toward parametric recall rather than random noise

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. [Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement](). In Findings of [EMNLP 2026](), October 2026, to appear.

Benchmarking context usage manipulation techniques

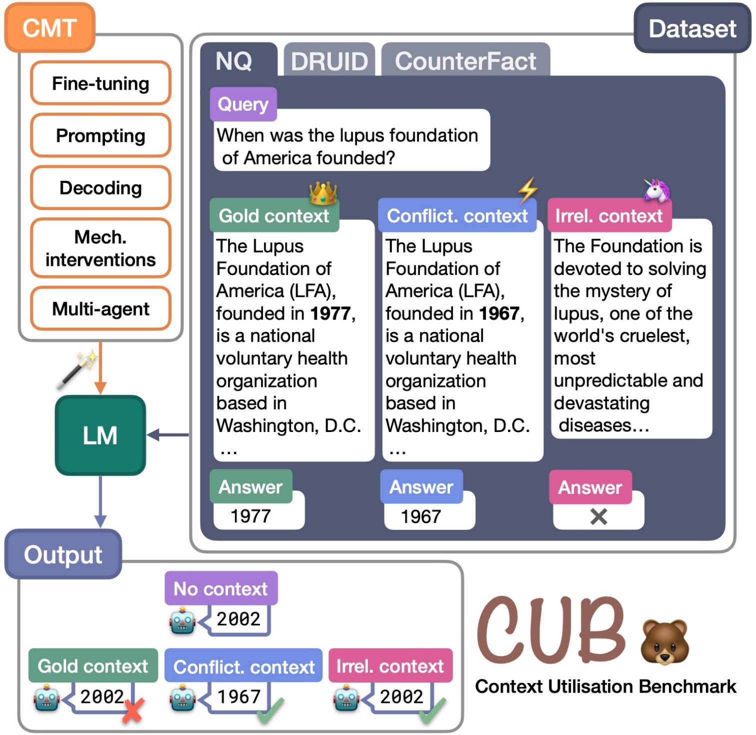
- Previous context usage experiments show that LLMs:
    - Struggle with more complex and long contexts
    - Can easily be distracted by irrelevant contexts due to context-memory conflicts

- 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. In Proceedings of ACL 2026, July 2026.

Benchmarking context usage manipulation techniques



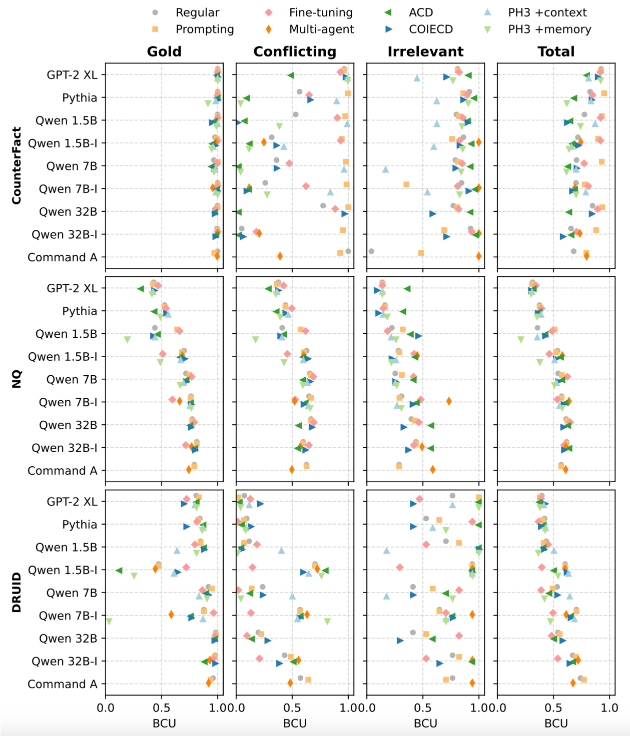
Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. [CUB: Benchmarking Context Utilisation Techniques for Language Models](). In Proceedings of [ACL 2026](), July 2026.

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. In Proceedings of ACL 2026, July 2026.

Are larger models better at utilising context?



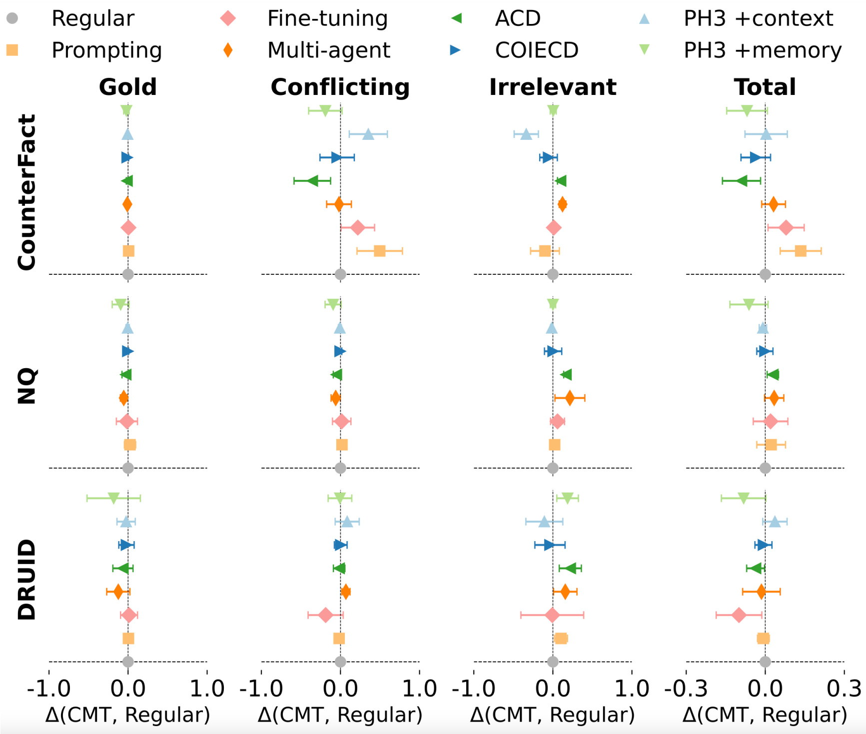
Binary context utilisation (BCU) score:

- 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 prediction (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]. In Proceedings of [ACL 2026], July 2026.

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](). In Proceedings of [ACL 2026](), July 2026.

Findings: 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. In Proceedings of ACL 2026, July 2026.

Answering RQs: Understanding LLMs’ Knowledge Utilisation

Do models **use context**?
✅ Yes, but utilisation is often fragile and conflict-dependent

What **characteristics** affect context usage?
➡️ Similarity, fluency, length (“easy to understand”)

When are those **sensitivities learned**?
➡️ Largely during instruction tuning, especially SFT

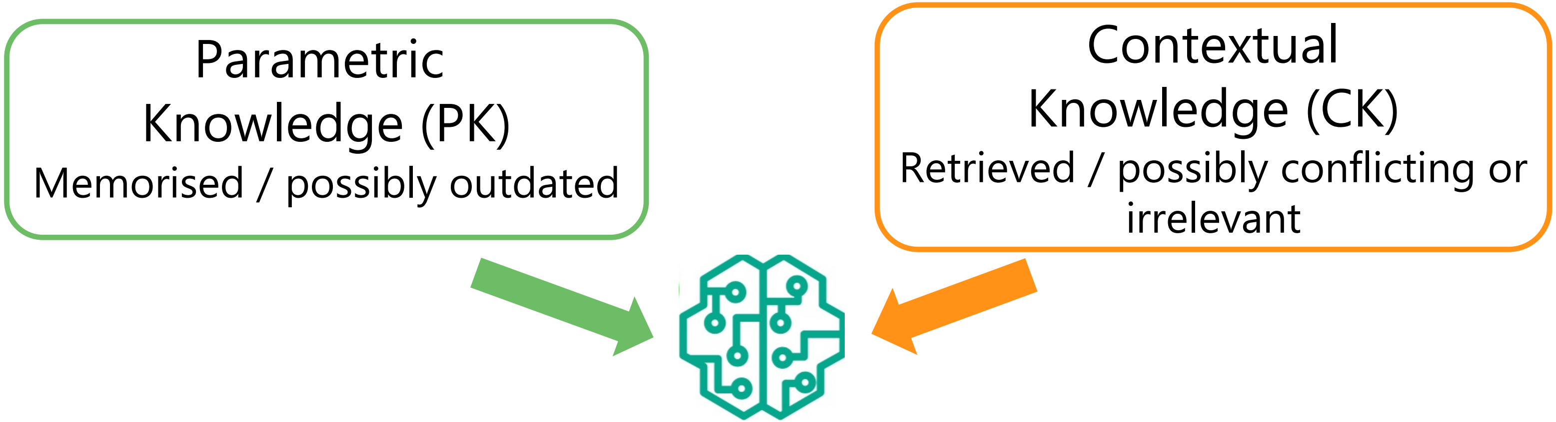
Can explanations tell us **which context** was used?
⚠️ Partially; they often identify that context mattered, not which evidence caused the prediction

Can we directly **analyse interactions** between parametric and contextual knowledge?
✅ Yes; interaction analyses reveal richer dynamics than a simple PK-versus-CK view

Can we **improve** context utilisation?
⚠️ Partially; existing techniques help, but no method is consistently best across all settings

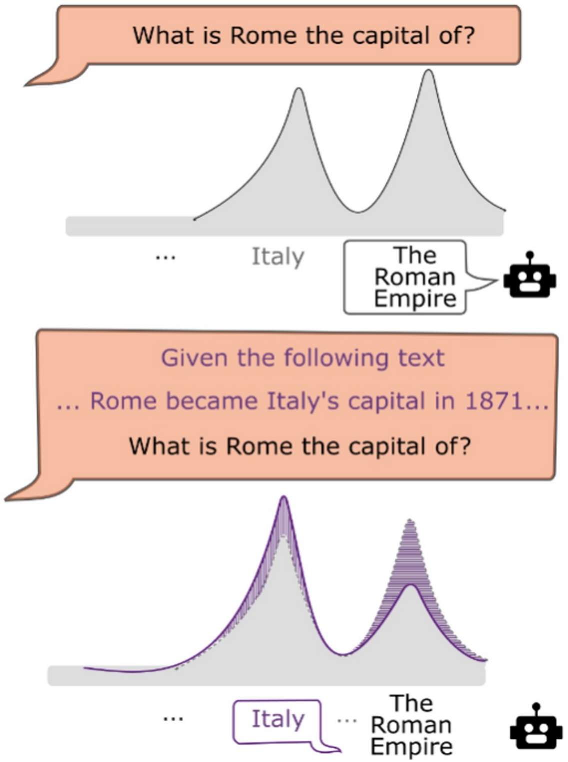
Wrap-Up: Utilisation of Knowledge by LLMs

- Context utilisation is not a fixed model capability
- It is shaped by conflicts, context characteristics, and instruction tuning
- Measuring context utilisation is itself challenging
- Realistic RAG settings expose weaknesses hidden by synthetic benchmarks
- Better analysis and intervention methods remain necessary



Why prevailing evaluation settings overestimates progress

- Context alignment ≠ causal dependence
- Synthetic evidence exaggerates both conflicts and success
- End-to-end accuracy evaluation hides failure modes of context usage



Implications for future research

- Scaling and longer contexts help, but do not resolve knowledge conflicts
- Evaluation must test counterfactual reliance on context
- RAG system design should account for when parametric knowledge should or should not be overridden

CopeNLU Lab

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

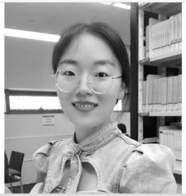
Pepa Atanasova
Assistant Professor
Pepa's research interests include the development, diagnostics, and application of explainability and interpretability techniques for NLP models.

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

Arnav Arora
PhD Student
Arnav's research interests include equitable ML, mitigating online harms, and the intersection of NLP 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 interest include multilingual models, fairne and accountability in NLP systems.

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.

Jean Seo
PhD Student
Jean's research interests include improving the safety of language models through explainability ar evaluation.

Philip Müller
PhD Student
Philip's main research interests include explainability, retrieval-augmented generation, and uncertainty quantification for LLM

Jeremy Herbst
PhD Student
Jeremy's main research interests include mechanistic interpretability, AI safety, and understanding the representations and mechanisms underlying AI models.

Lenka Tětková
Postdoc
Lenka is a DDSA Postdoctoral Fellow at the Technical University of Denmark, working on concept-based explainability and human-machine alignment, and collaborates with Pepa on her project on the geometry of trust in latent representations.

Bao Dinh
Visiting Student
Bao Dinh is a visiting master's student from MBZUAI, UAE. His research focuses on adversarial attacks on aligned LLMs, with broader interests in AI safety and representation alignment.

Syed Ali Redha Alsagoff
Visiting Student
Ali is a visiting bachelor's student from Nanyang Technological University, Singapore. His research focuses on interpretability and reasoning in LLMs.

+ You?
We're hiring! ➡️

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 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.

Nadya Yuki Wangsajaya, Haeun Yu, **Isabelle Augenstein**. Emergence of Context Characteristics Sensitivity in Large Language Models. CoRR, abs/2606.09525, June 2026.

Jingyi Sun*, Pepa Atanasova*, Sagnik Ray Choudhury, Sekh Mainul Islam, **Isabelle Augenstein**. Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models. Computational Linguistics, August 2026.

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. In Findings of EMNLP 2026, October 2026, to appear.

Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. In Proceedings of ACL 2026, July 2026.

COPENLU

NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

Thank you for your attention! Questions?

We're hiring! ⬇️