# Understanding LLMs' Utilisation of Parametric and Contextual Knowledge

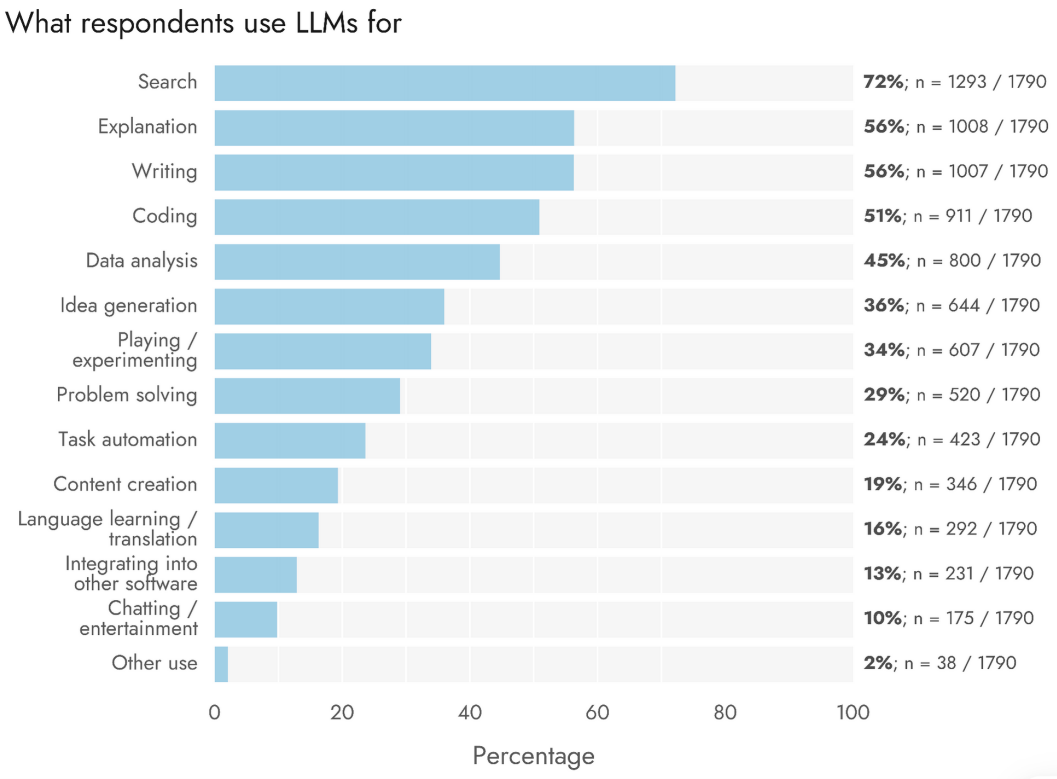
**Isabelle Augenstein**

ALPS
31 March 2026

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

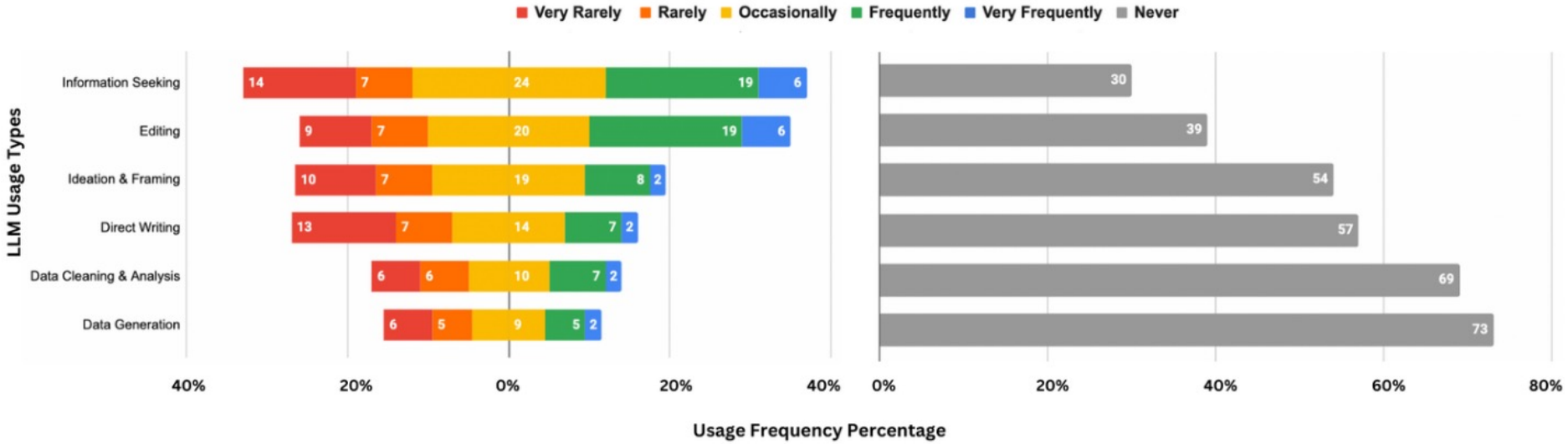
UNIVERSITY OF COPENHAGEN

# Usage of Large Language Models



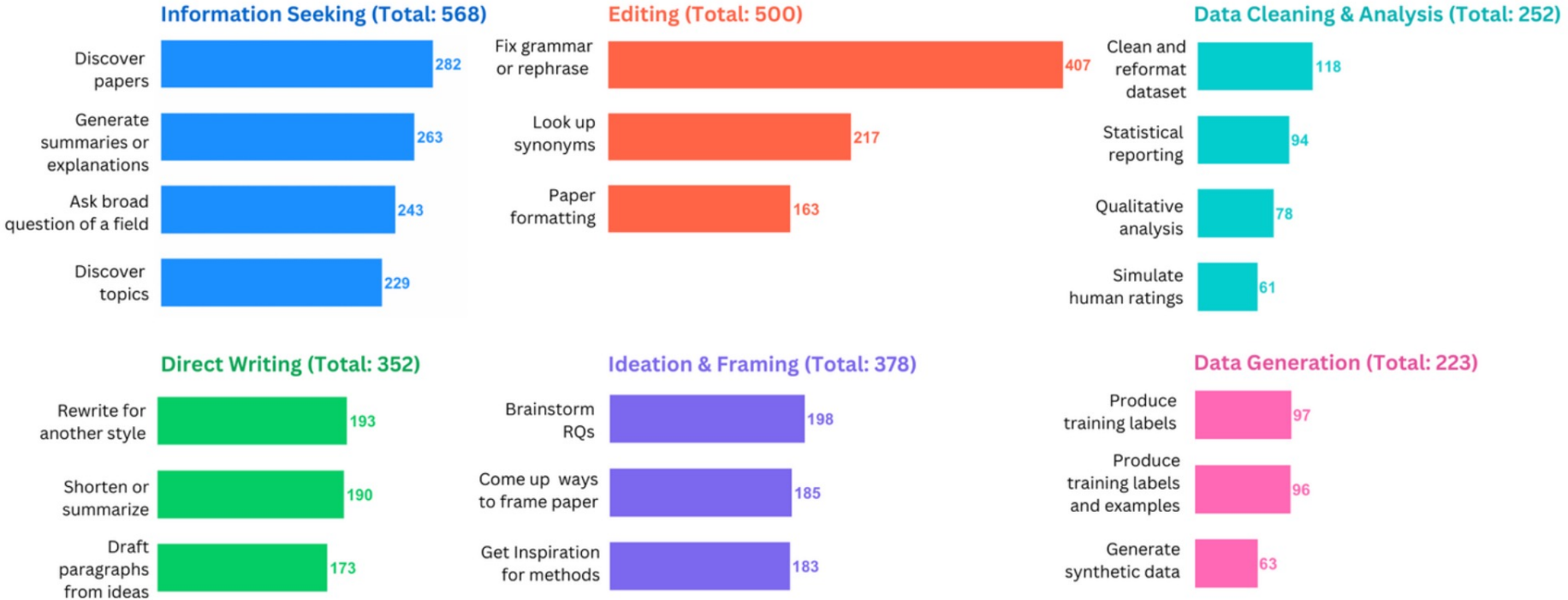

Adoption and uses of LLMs among US tech workers (Rethink Priorities, 2025). https://rethinkpriorities.org/research-area/adoption-llms-tech-workers/

# Usage of Large Language Models



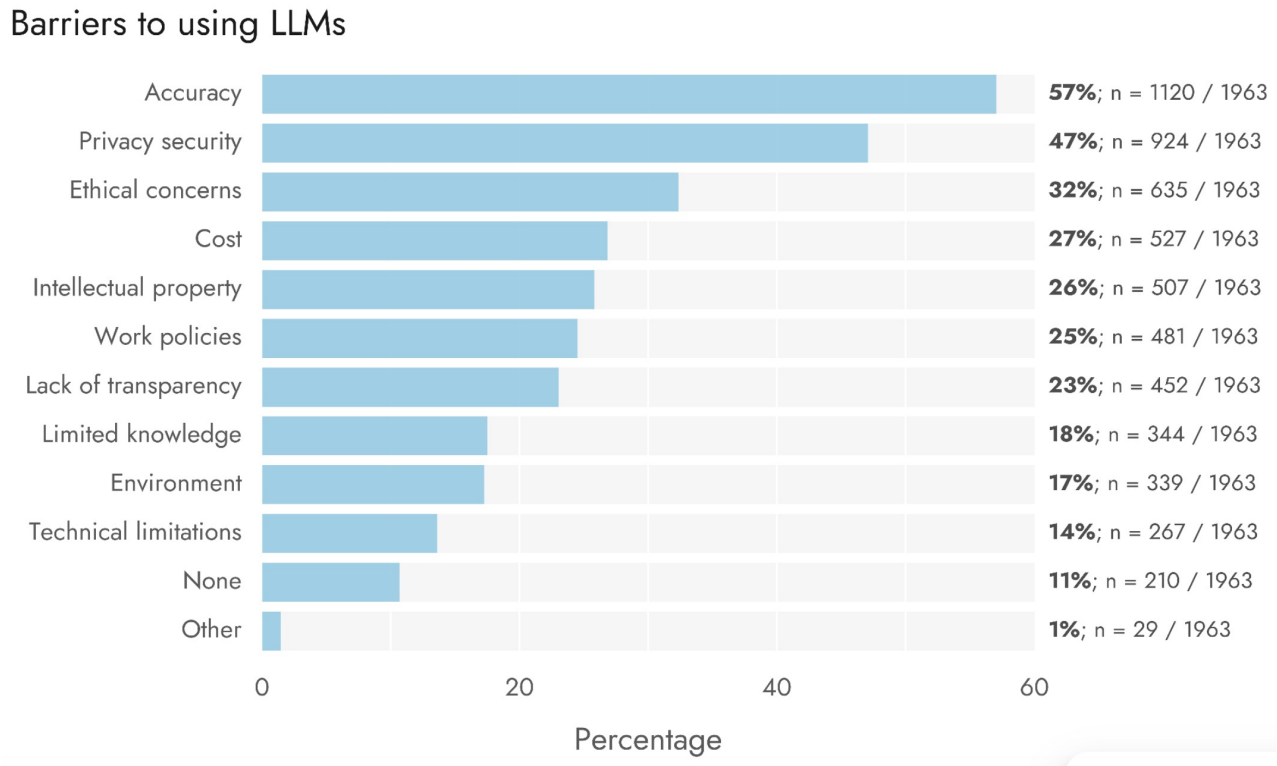

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. In Proceedings of COLM, 2025.

# Usage of Large Language Models



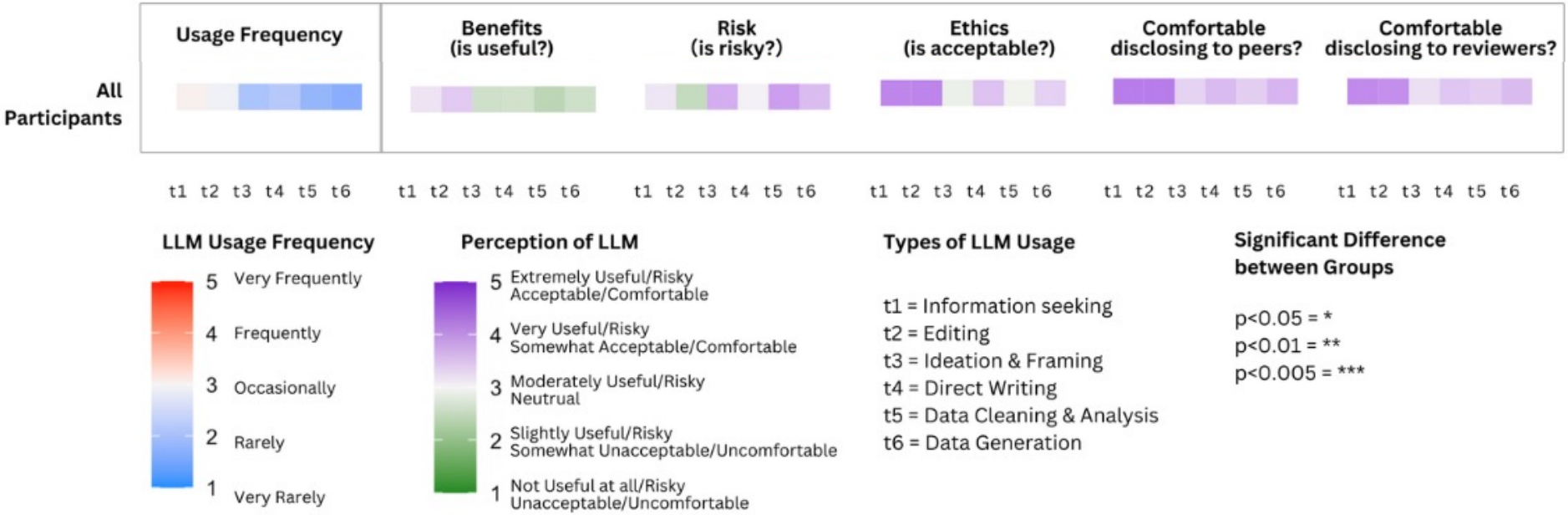

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. In Proceedings of COLM, 2025.

# Risks of Using of Large Language Models




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

# LLM Usages – Benefits vs Risks



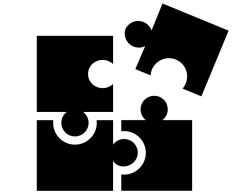

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. In Proceedings of COLM, 2025.

# LLM Usages – Benefits vs Risks

| Theme | Description | Example |
|---|---|---|
| **Hallucination & Misinformation** | Production and spread of incorrect information invented by the model | *"Sometimes it creates so complicated hallucinations so that even an expert can think that what it writes it true although it is not."*<br>*"Putting more falsehoods into [the internet's] shared memory is a crime."* |
| **Inaccuracy** | Incorrect conclusions and analyses | *"There is a risk of less experienced scientists using these technologies as they are unable to check if the outputs are correct as easily as someone with more experience/intuition."*<br>*"The risks are proportional to prior knowledge of the subject."* |
| **Fabrication** | Using LLMs to fabricate data and research results | *"The risk of reporting 'results' based on synthetic data without actually having conducted any experiment."*<br>*"LLMs are tools for automated plagiarism and data fabrication that pose an existential threat to the network of trust essential for the integrity of academic work and the proper attribution of credit."* |

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. In Proceedings of COLM, 2025.

# Factuality Challenges of Large Language Models

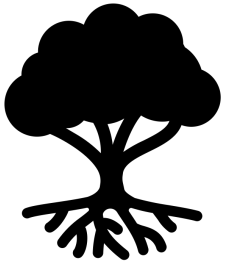


Citation Gaps

Truthfulness

Fluent Style

Outdated Knowledge



Grounding Deficiency

Confident Tone

Halo Effect

Unreliable Evaluation

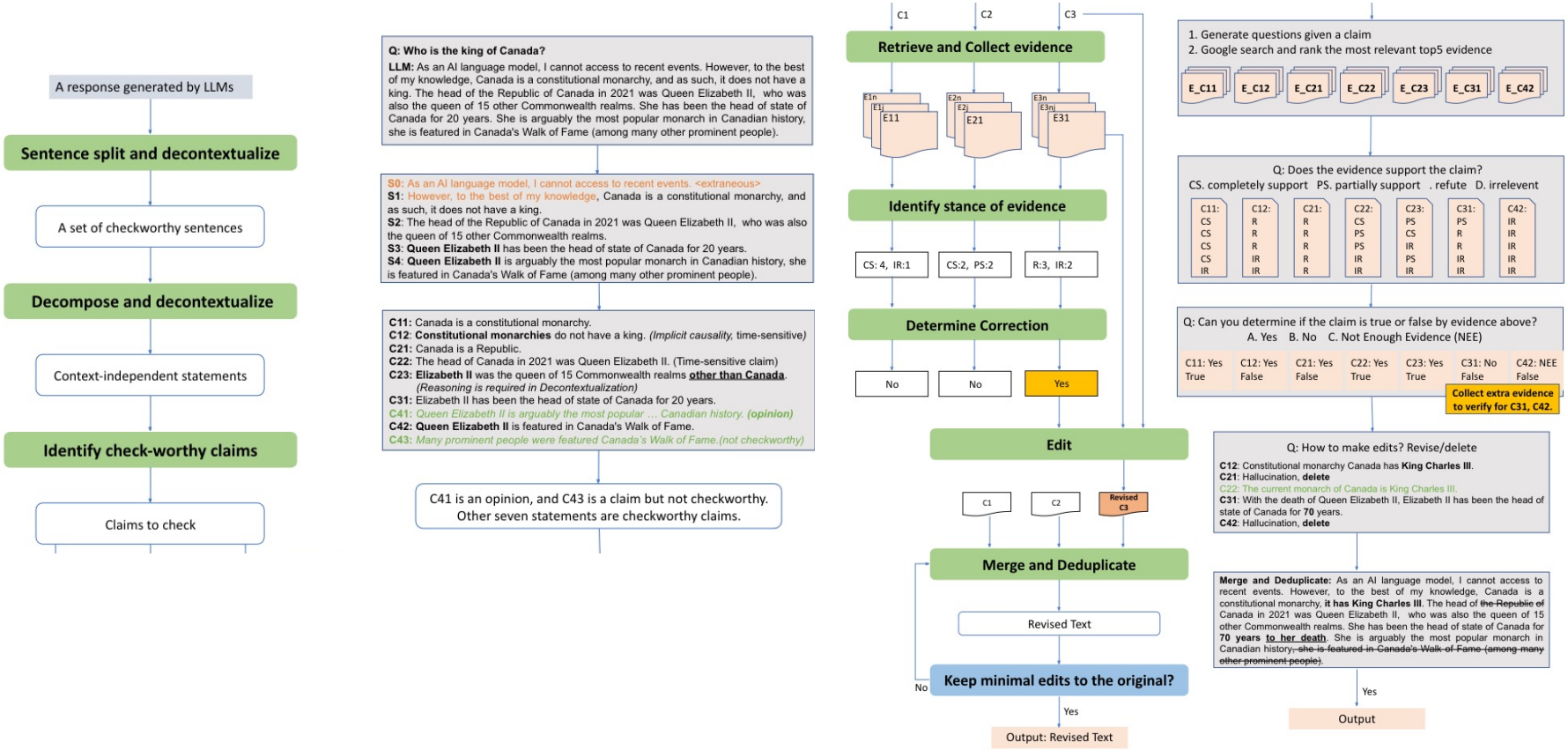
Augenstein et al. (2024). Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.

# Factuality Challenges in the Era of LLMs

- Addressing threats:
  - Making LLMs safer – data cleansing, watermarking, privacy etc.
  - Modularised knowledge-grounded framework
  - **Retrieval-augmented generation**
  - **Detecting and correcting factual mistakes** at inference time
  - **Better evaluation**
  - Recognising AI-generated content
  - AI regulation
  - Public education

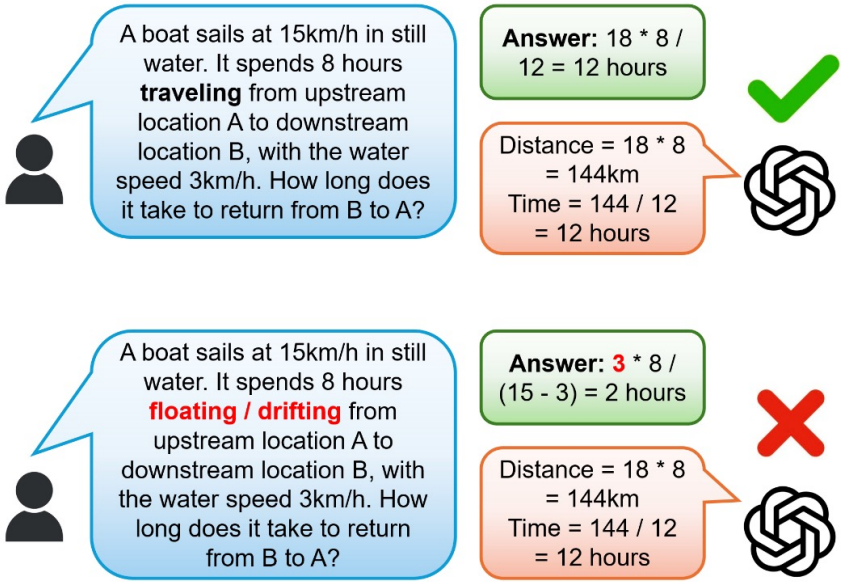
Augenstein et al. Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.

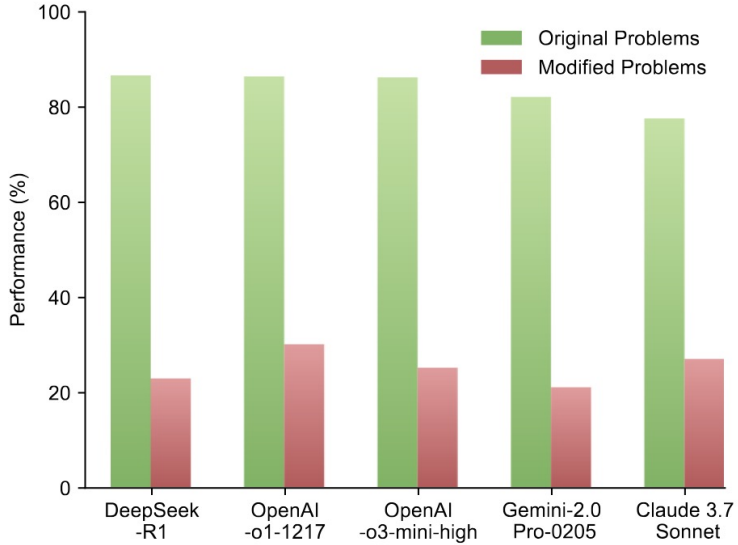
# Fact Checking of Machine-Generated Misinformation




Wang et al. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In EMNLP 2024, November 2024.

# LLM Prompt Instability

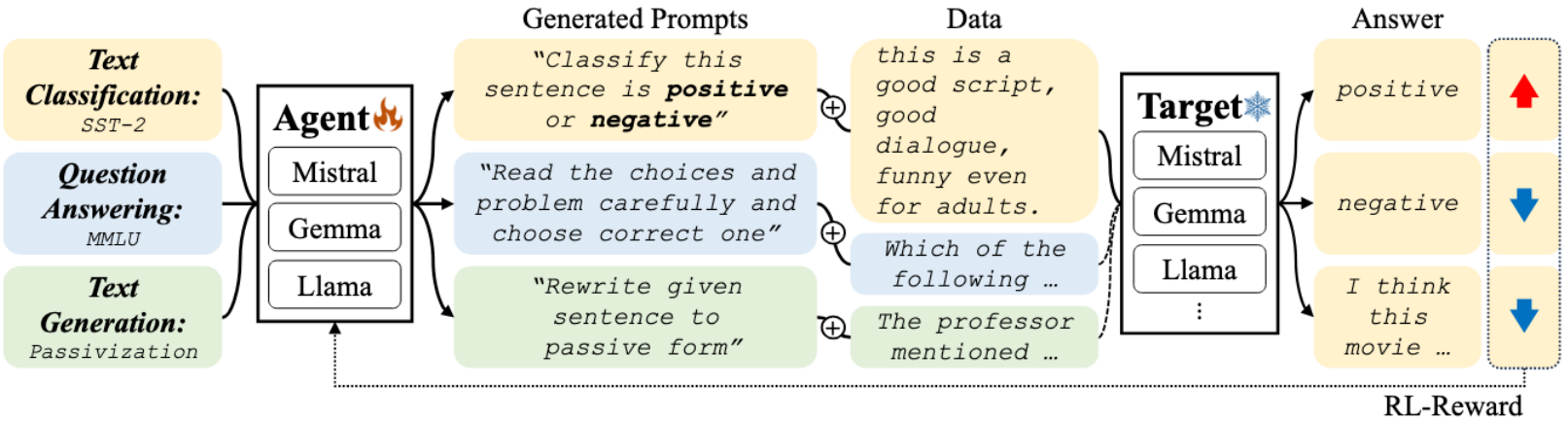



Drastic performance drops when performing small alterations to wording

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.

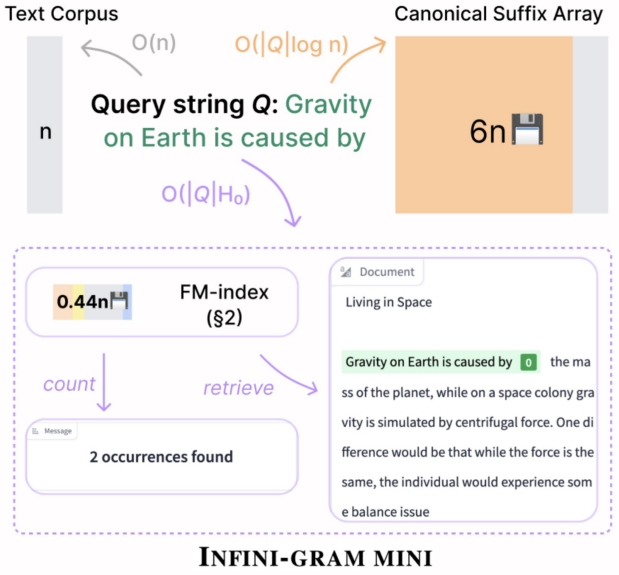
# LLM Prompt Instability -> Prompt Tuning




StablePrompt. We formulate prompt tuning as an RL-framework using LLMs. We use the target LLM and the given dataset as the world model, and the agent LLM as the policy. We use the response of the target LLM to the prompt generated by the agent LLM as the reward

Kwon et al. (2024). StablePrompt : Automatic Prompt Tuning using Reinforcement Learning for Large Language Model. In EMNLP.

# Evaluation of Benchmark Contamination

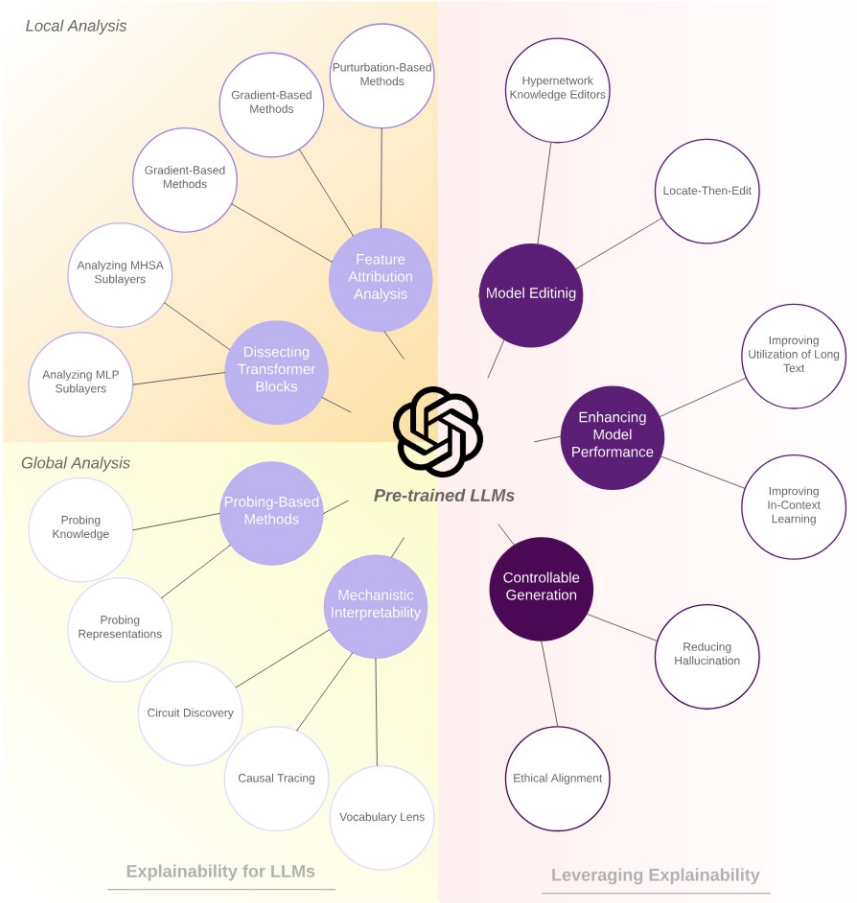



| | Test Size | Pile train | DCLM baseline | CC 2025-05 | CC 2025-08 | CC 2025-13 | CC 2025-18 | CC 2025-21 | CC 2025-26 |
|---|---|---|---|---|---|---|---|---|---|
| **Knowledge and Reasoning** | | | | | | | | | |
| MMLU | 1000 | 13.20 | 28.40 | 13.50 | 9.00 | 12.10 | 11.50 | 11.70 | 9.20 |
| MMLU-Pro | 1000 | 5.50 | 16.20 | 7.10 | 5.40 | 6.00 | 6.30 | 7.40 | 6.90 |
| BigBenchHard | 1000 | 0.00 | 0.10 | 1.40 | 1.40 | 3.20 | 2.30 | 1.80 | 1.70 |
| AGIEval | 1000 | 0.80 | 3.10 | 2.70 | 3.60 | 3.00 | 7.00 | 9.40 | 4.60 |
| GPQA | 448 | 0.00 | 0.00 | 0.90 | 2.00 | 1.30 | 0.70 | 0.90 | 2.70 |
| HLE | 881 | 0.00 | 0.30 | 0.10 | 0.00 | 0.10 | 0.00 | 0.00 | 0.00 |
| **Math** | | | | | | | | | |
| AIME-2024 | 30 | 0.00 | 0.00 | 10.00 | 3.30 | 6.70 | 40.00 | 40.00 | 13.30 |
| GSM8K | 1000 | 0.00 | 5.00 | 5.00 | 0.80 | 6.90 | 0.70 | 74.20 | 7.30 |
| MATH-500 | 500 | 0.60 | 3.20 | 0.60 | 7.80 | 0.80 | 0.80 | 0.80 | 8.20 |
| MGSM | 250 | 0.00 | 0.00 | 5.60 | 1.60 | 35.60 | 0.80 | 72.80 | 6.00 |
| **Code** | | | | | | | | | |
| HumanEval | 164 | 0.00 | 0.00 | 0.00 | 0.60 | 0.60 | 0.60 | 0.00 | 0.00 |
| HumanEval+ | 164 | 0.00 | 0.00 | 0.00 | 0.60 | 0.60 | 0.60 | 0.00 | 0.00 |
| LiveCodeBench | 880 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SWE-bench | 500 | 0.00 | 0.00 | 0.20 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 |
| MBPP | 500 | 0.00 | 0.40 | 1.00 | 1.40 | 1.20 | 1.80 | 1.00 | 1.40 |
| **Commonsense Understanding** | | | | | | | | | |
| ARC-Challenge | 1000 | 1.80 | 34.10 | 11.90 | 4.00 | 3.10 | 3.80 | 4.20 | 4.80 |
| ARC-Easy | 1000 | 1.30 | 31.70 | 5.40 | 9.50 | 5.50 | 5.50 | 6.10 | 6.20 |
| CSQA | 1000 | 0.10 | 1.00 | 0.10 | 0.10 | 0.20 | 0.10 | 0.00 | 0.10 |
| HellaSwag | 1000 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 |
| OpenbookQA | 500 | 10.80 | 15.60 | 14.60 | 30.20 | 13.20 | 13.40 | 13.20 | 12.20 |
| Social IQa | 1000 | 0.00 | 0.50 | 0.20 | 4.40 | 0.20 | 0.30 | 0.20 | 0.10 |
| WinoGrande | 1000 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Reading Comprehension** | | | | | | | | | |
| CoQA | 500 | 8.00 | 18.40 | 7.40 | 8.80 | 8.60 | 7.20 | 7.60 | 8.80 |
| SQuAD | 1000 | 2.80 | 40.10 | 2.70 | 33.00 | 10.10 | 1.50 | 2.00 | 8.50 |

Efficient search over LLM pre-training data, reveals heavy **benchmark contamination**

Xu et al. (2025). Infini-gram mini: Exact n-gram Search at the Internet Scale with FM-Index. EMNLP 2025, **best paper**.

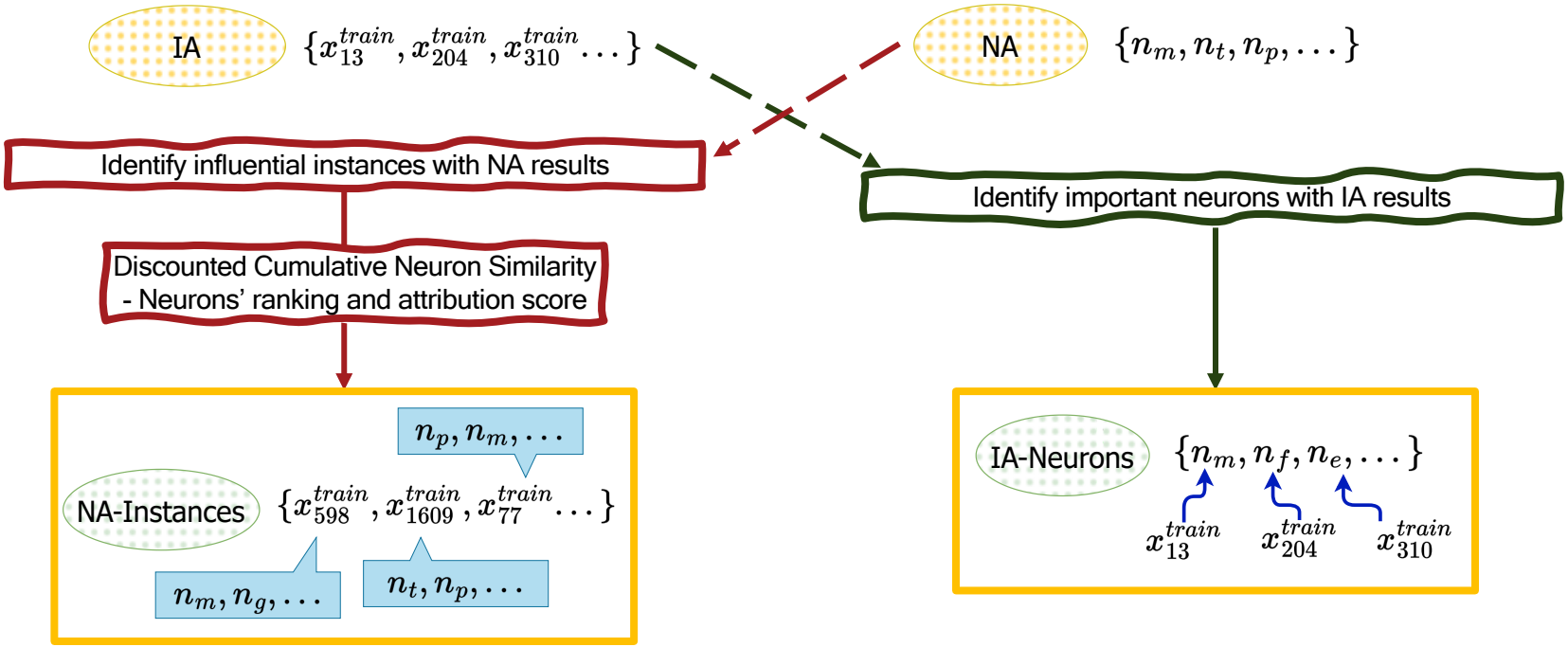
# Interpretability




Luo & Specia (2024). From Understanding to Utilization: A Survey on Explainability for Large Language Models. Arxiv 2401.12874.

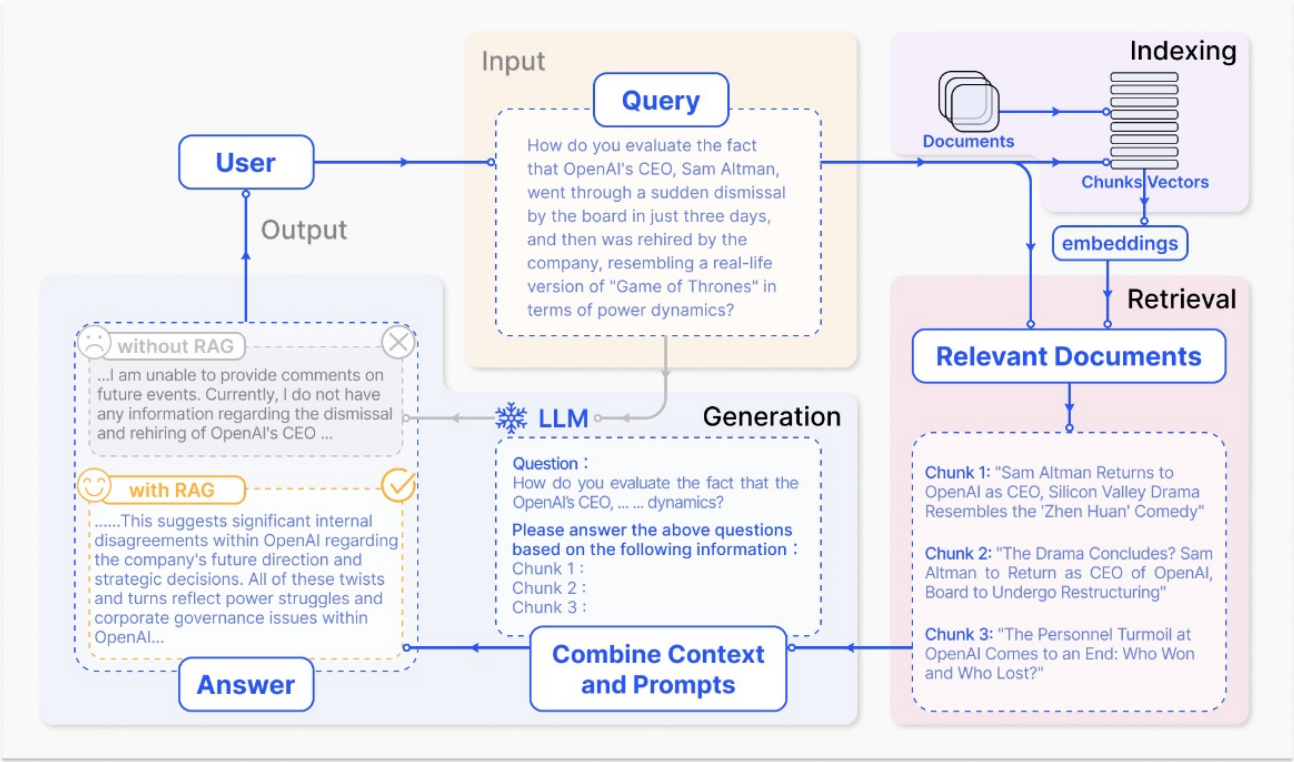
# An Evaluation Framework for Attribution Methods

## 1) Aligning the Results of Attribution Methods



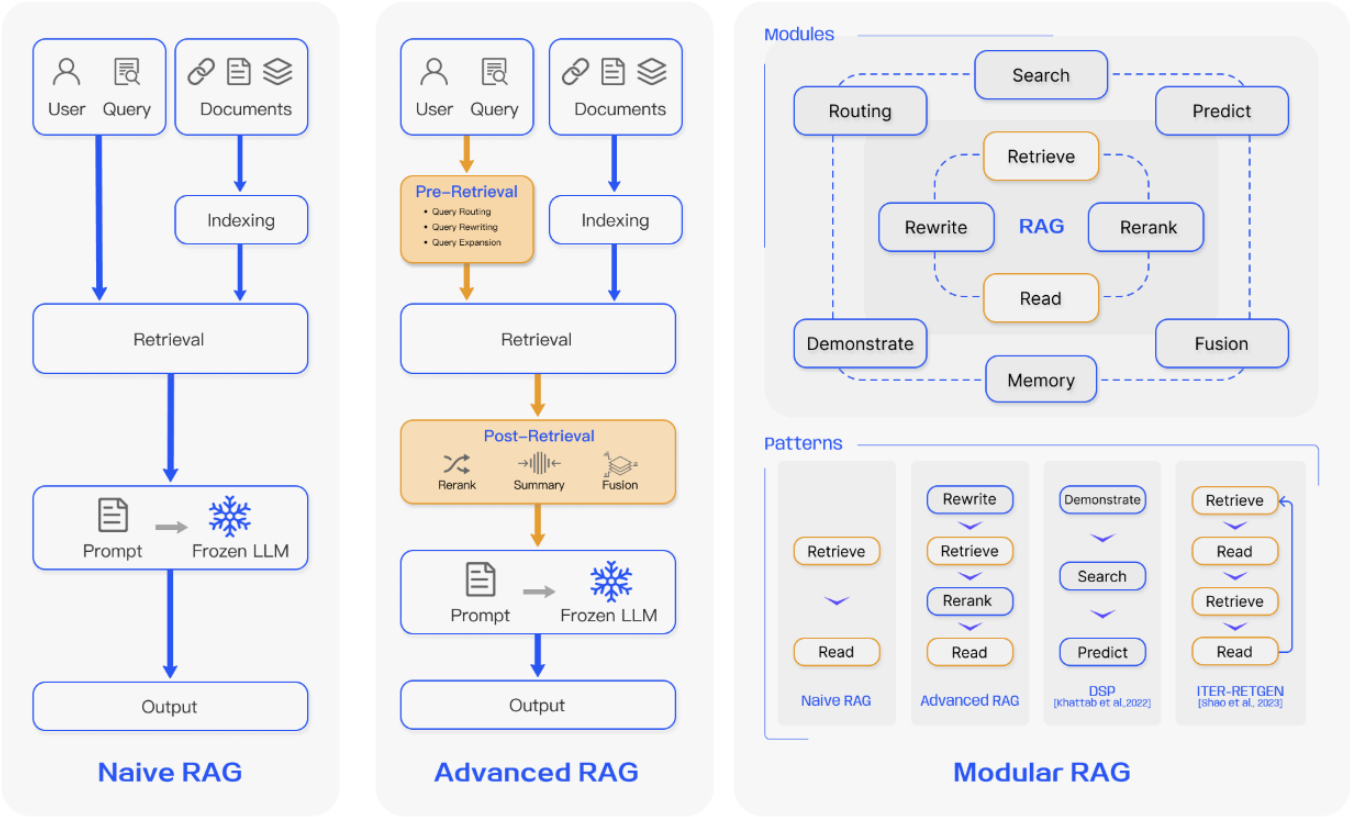

Haeun Yu, Pepa Atanasova, **Isabelle Augenstein**. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), August 2024.

# Augmentation of LLMs with External Knowledge



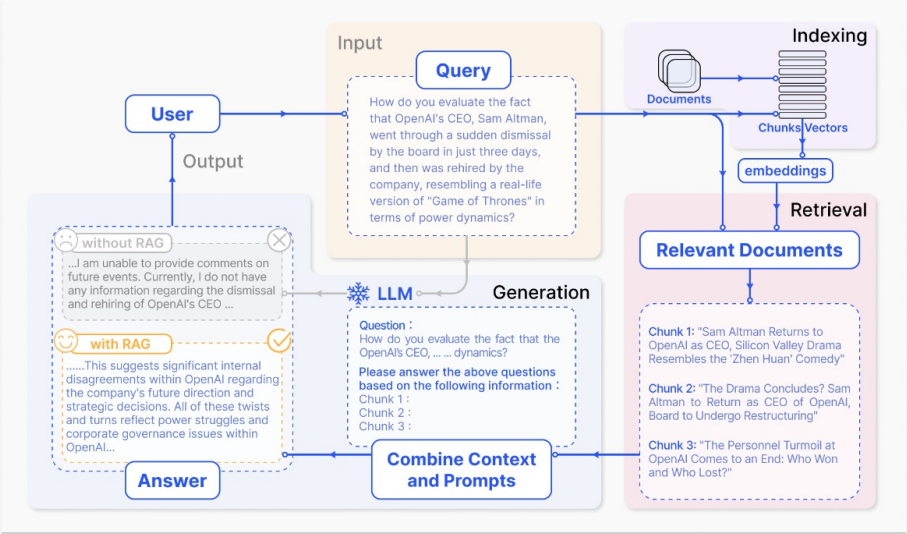

Gao et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arxiv:2312.10997.

# Retrieval-Augmented Generation




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

# Augmentation of LLMs with External Knowledge

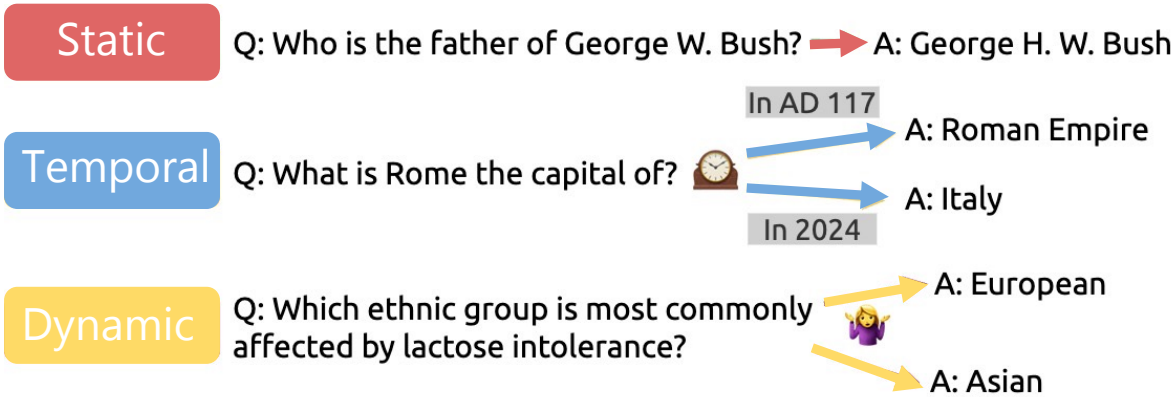



- *Retrieving contextual knowledge to augment LLM's parametric knowledge*
- *Can better take context-dependent nature of queries into account*
- *Interplay between contextual and parametric knowledge underexplored*
- *When should contextual knowledge overwrite or augment parametric knowledge?*

# 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
  - Explaining context usage of RAG models
  - 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*.

# Wikipedia Controversial Topics

https://en.wikipedia.org/wiki/Category:Wikipedia_controversial_topics 120%

## Pages in category "Wikipedia controversial topics"

The following 200 pages are in this category, out of approximately 3,909 total. This list may not reflect recent changes.

(previous page) (next page)

- Wikipedia:List of controversial issues

.

- Talk:.eco

*

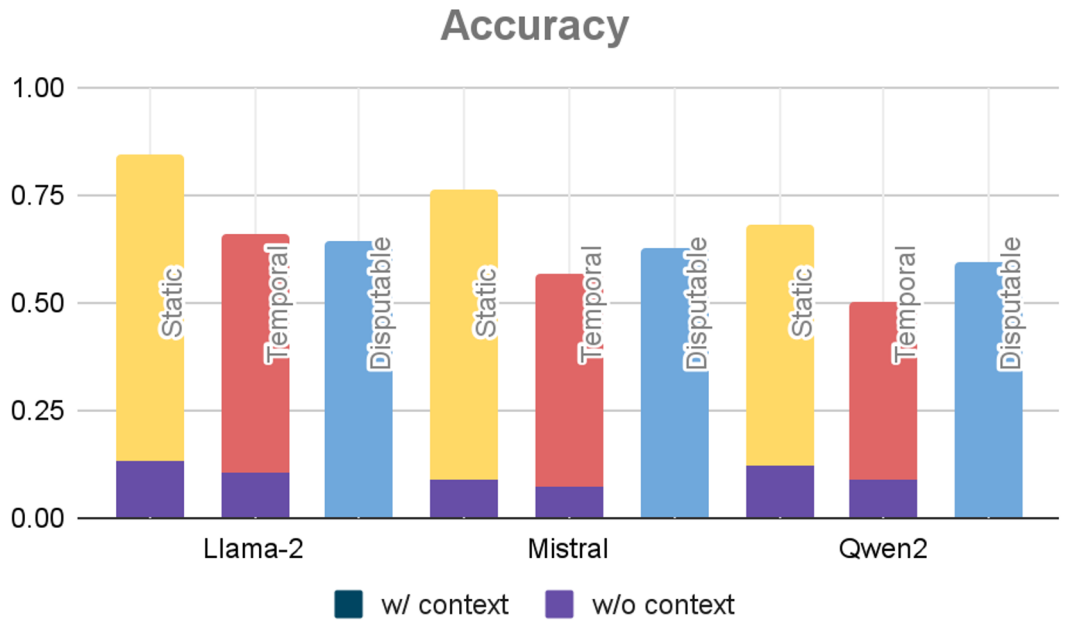
- Wikipedia:Controversial articles

### 0–9

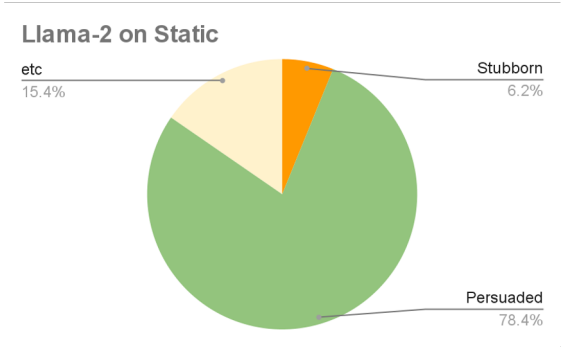
- Talk:2G spectrum case
- Talk:4B movement
- Talk:4chan
- Talk:4chan/Archive 16
- Talk:6ix9ine
- Talk:7 World Trade Center
- Talk:8chan
- Talk:9/11 conspiracy theories
- Talk:9/11 conspiracy theories regarding Jews or Israel
- Talk:10/40 window
- Talk:12 May Karachi riots
- Talk:40 Days for Life
- Talk:44M Lidérc
- Talk:50 Cent Party
- Talk:123Movies
- Talk:420chan
- Talk:1421: The Year China Discovered the World
- Talk:2009 Iranian presidential election
- Talk:2009 Mangalore pub attack
- Talk:2010–2012 Algerian protests
- Talk:2011 Alexandria bombing
- Talk:2011 England riots
- Talk:2011 Rome demonstration
- Talk:2011 Super Outbreak/Archive 3
- Talk:2011–2012 Iranian protests
- Talk:2011–2012 Moroccan protests
- Talk:2012
- Talk:2012 anti-Japanese demonstrations in China
- Talk:2012 Aurora theater shooting
- Talk:2012 phenomenon
- Talk:2012 United Nations Climate Change Conference
- Talk:2013 Egyptian coup d'état
- Talk:2013 Mayflower oil spill
- Talk:2013 Muzaffarnagar riots
- Talk:2013 Neo Irakleio Golden Dawn office shooting
- Talk:2014 Crimean status referendum
- Talk:2014 Euromaidan regional state administration occupations
- Talk:2014 Oso landslide
- Talk:2014 pro-Russian unrest in Ukraine
- Talk:2015 Chapel Hill shooting
- Talk:2015 Ecuadorian protests
- Talk:2015–2016 protests in Brazil
- Talk:2016 Indian banknote demonetisation
- Talk:2021 United States Electoral College vote count
- Talk:2021 West Bengal post-poll violence
- Talk:2022 Al-Aqsa clashes
- Talk:2022 California Proposition 1
- Talk:2022 FIFA World Cup
- Talk:2022 Muhammad remarks controversy
- Talk:2022 West Bengal School Service Commission recruitment scam
- Talk:2022–2023 Pentagon document leaks
- Talk:2023 Indian wrestlers' protest
- Talk:2023 Kaveri water dispute protests
- Talk:2023 West Bengal local elections violence
- Talk:2023–2024 Gaza Strip preterm births
- Talk:2024 Ayta al-Shaab clashes
- Talk:2024 Azad Kashmir protests
- Talk:2024 Beqaa Valley airstrikes
- Talk:2024 constitutional reform attempts in the Philippines
- Talk:2024 Derdghaya Melkite Church airstrike
- Talk:2024 drone attack on Benjamin Netanyahu's residence
- Talk:2024 Hadera stabbing
- Talk:2024 Hezbollah drone strike on Binyamina
- Talk:2024 Indian farmers' protest
- Talk:2024 Iranian presidential election
- Talk:2024 Israeli invasion of Lebanon
- Talk:2024 Kafr Kila clashes

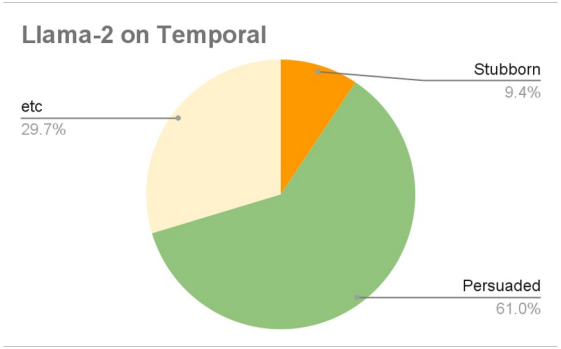
# How do LMs perform on the dataset?

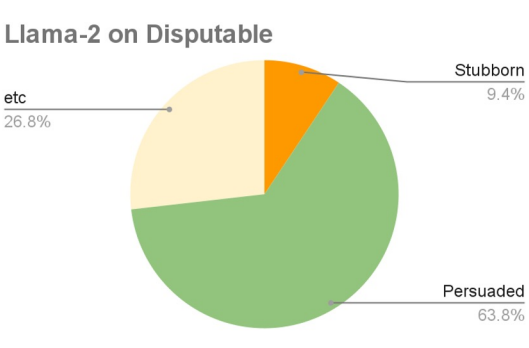
Models perform **best** on static questions, **with and without context.**



# How do LMs perform on the dataset?

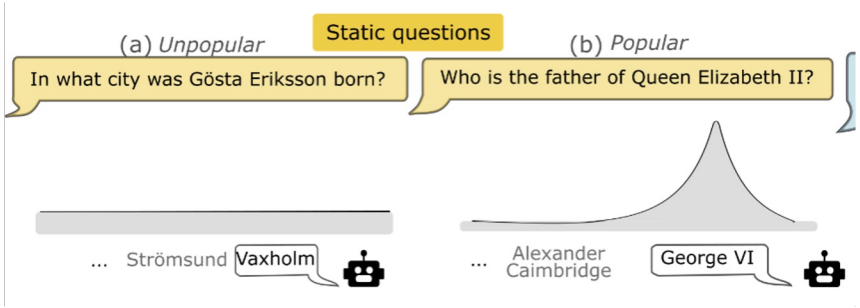


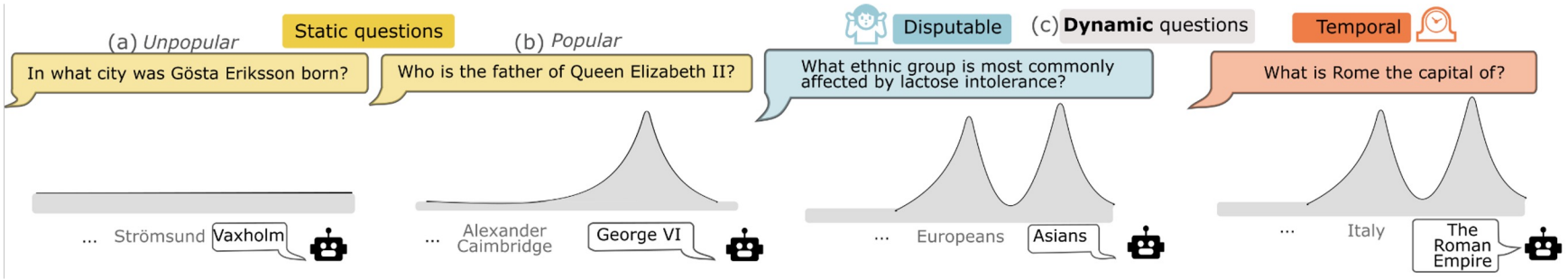




We see more **stubborn instances** in the dynamic partitions
-> Why are **dynamic** facts so **stubborn**?

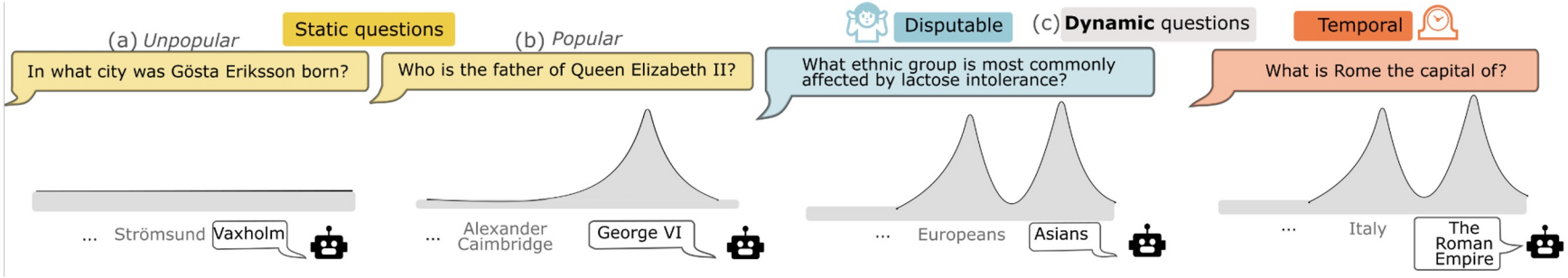
# Intra-Memory Conflict in Output Distribution



# Intra-Memory Conflict in Output Distribution



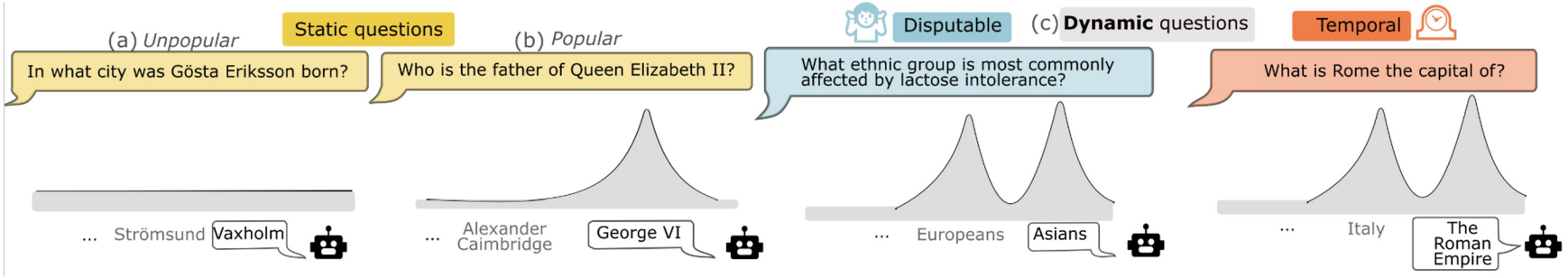
# Intra-Memory Conflict in Output Distribution




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

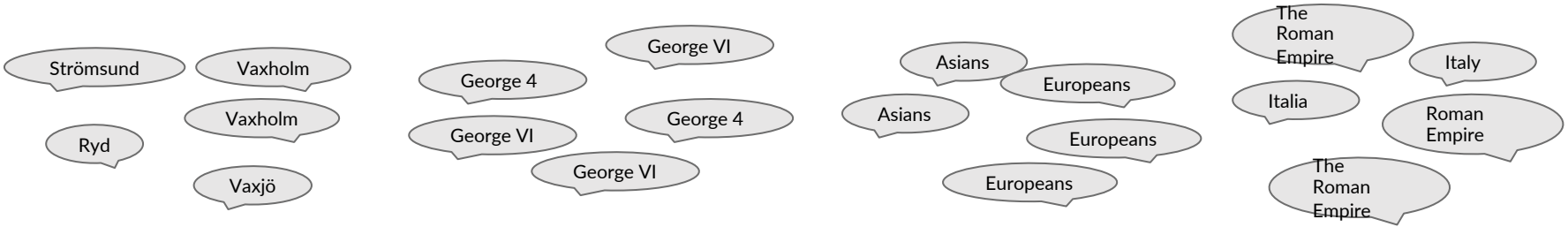
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*

# Intra-Memory Conflict in Output Distribution

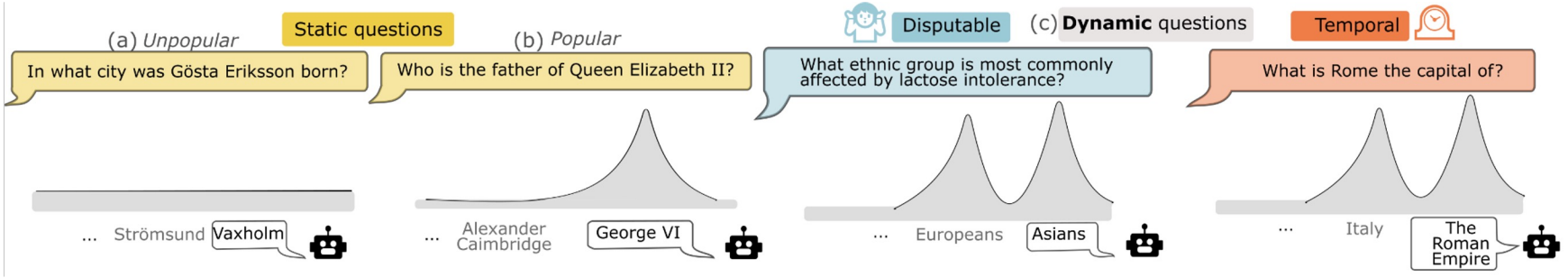



***Dynamic* facts should show greater *entropy* across objects.**

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

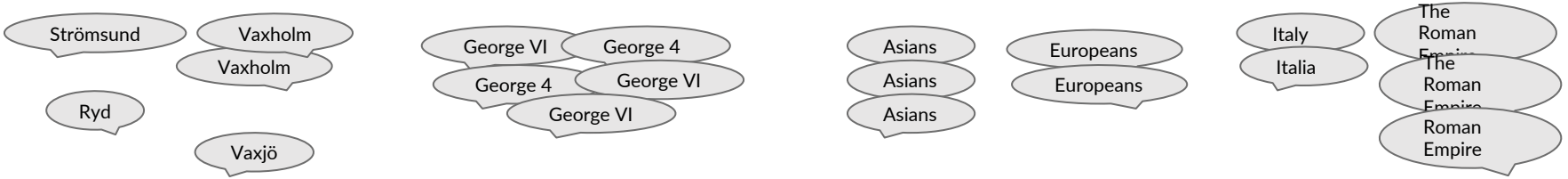


# Intra-Memory Conflict in Output Distribution

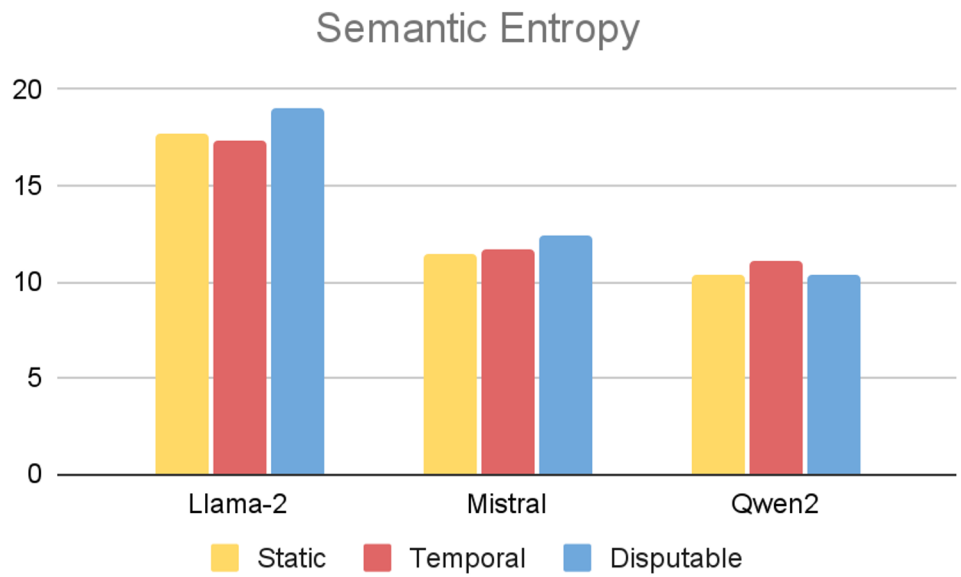



***Dynamic* facts should show greater *entropy* across objects.**

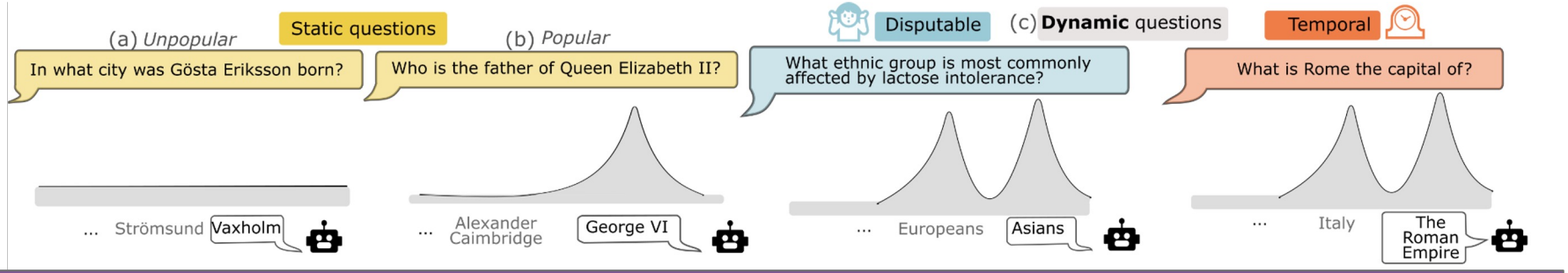
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*



# However, this is not always the case

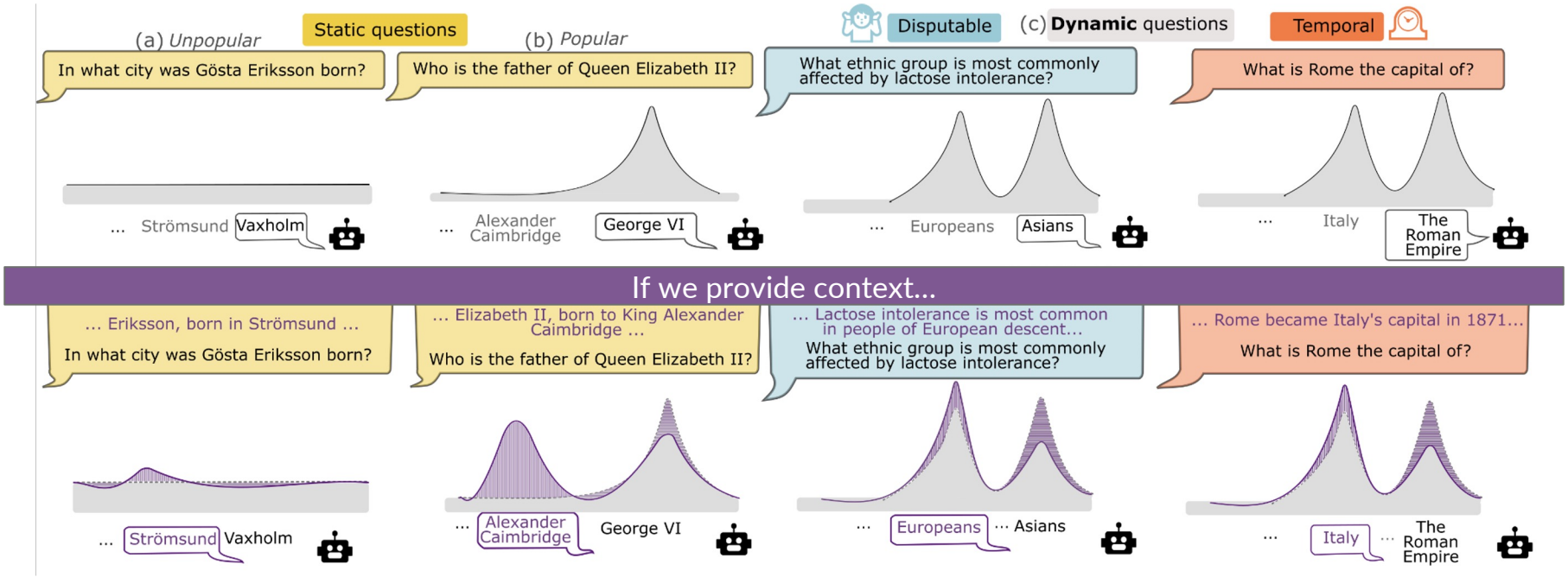


# Context-Memory Conflict

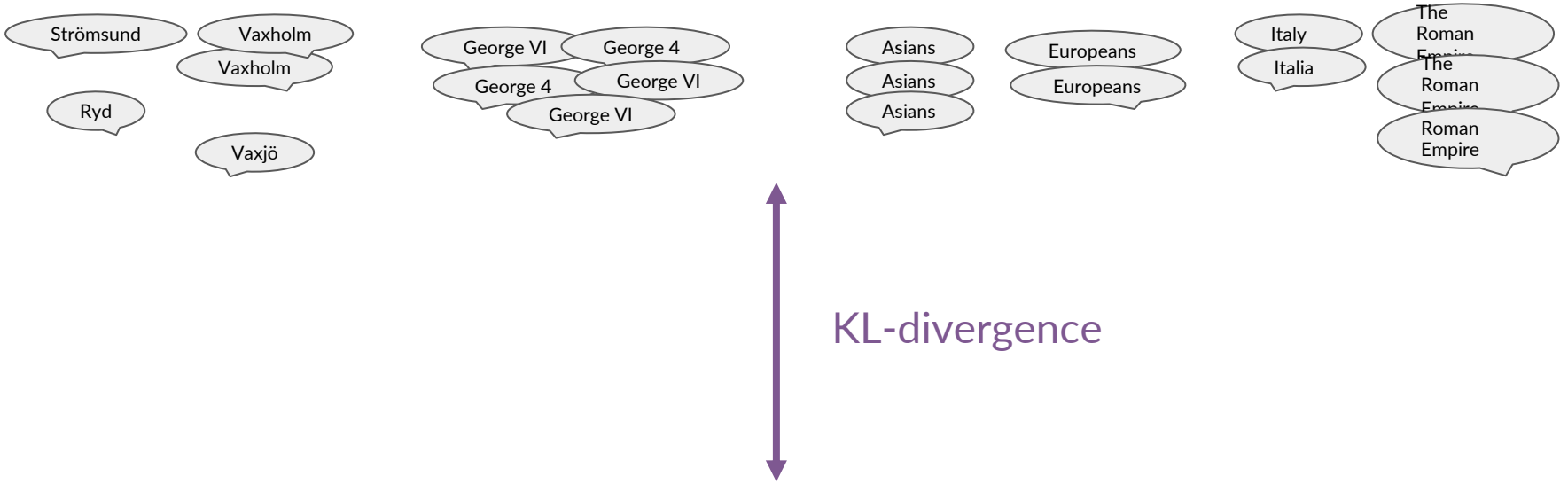



If we provide context...

# Context-Memory Conflict

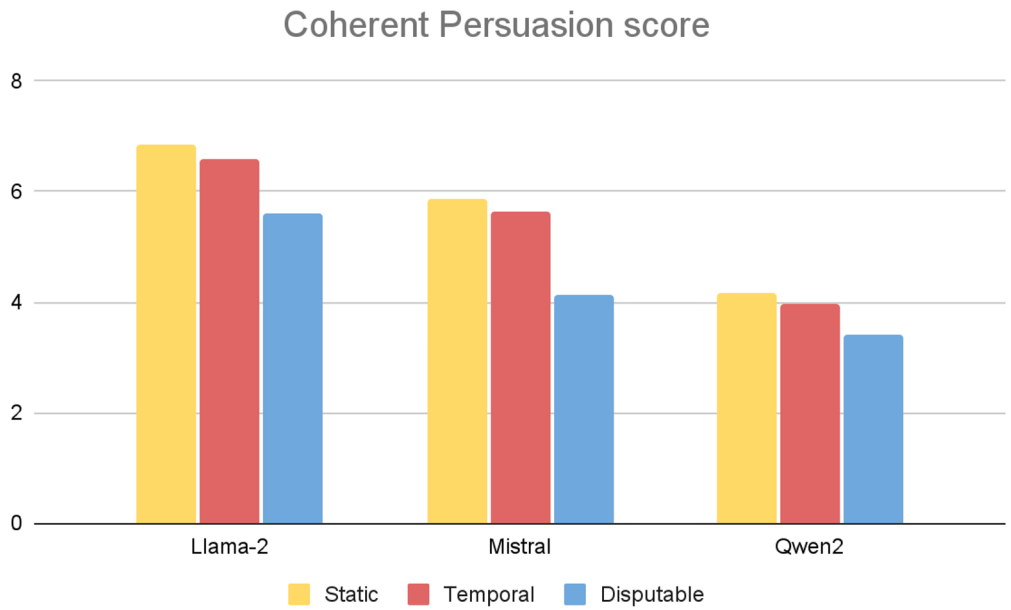


# Coherent Persuasion Score



# Persuasion Score across Partitions

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



# Persuasion Score across Partitions

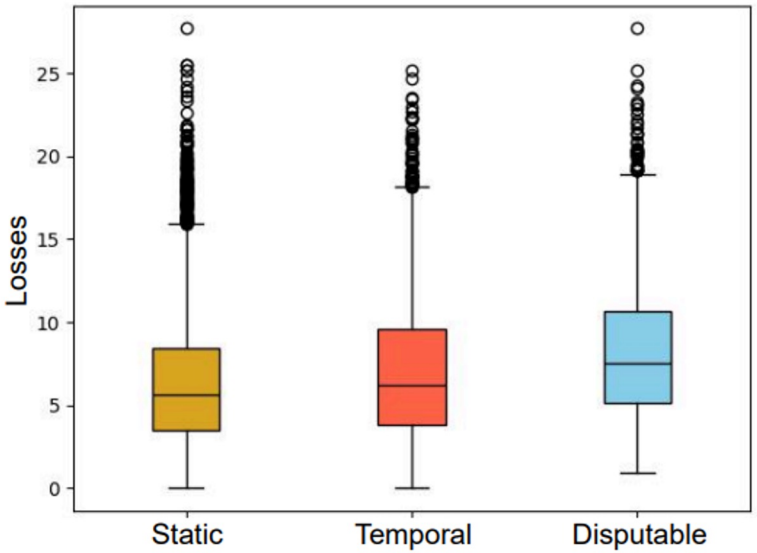
We see the **greatest persuasion score** for the **static dataset.**

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.

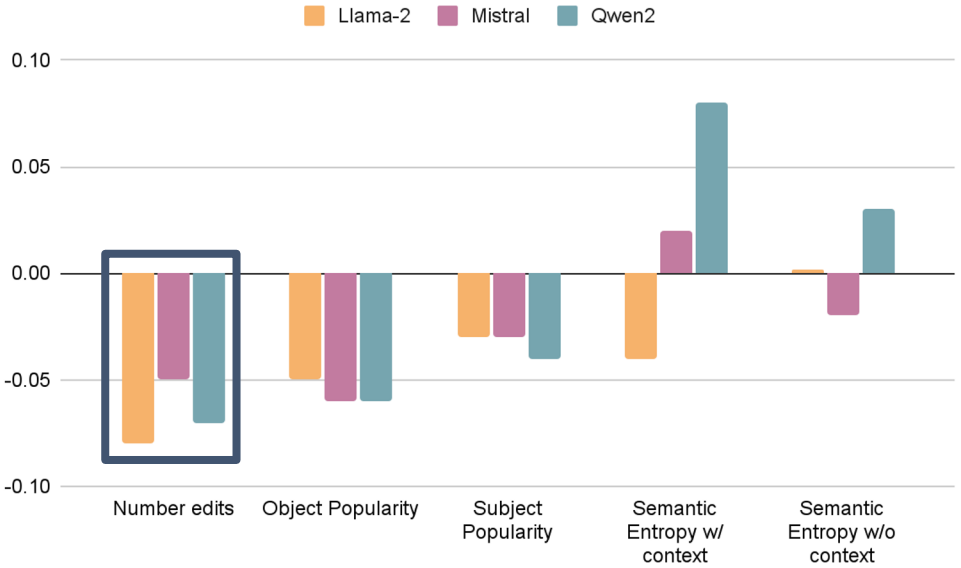
A higher loss indicates greater change required to steer the LM to output the target answer.

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

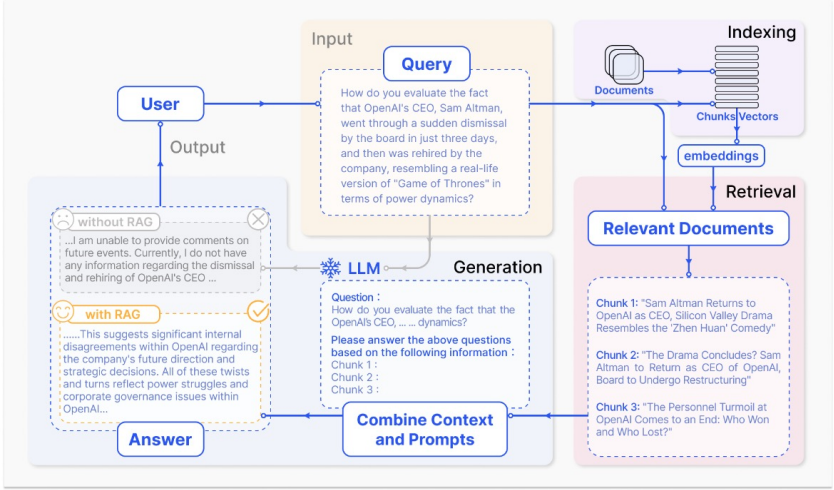
Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

# 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
  - Explaining context usage of RAG models
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

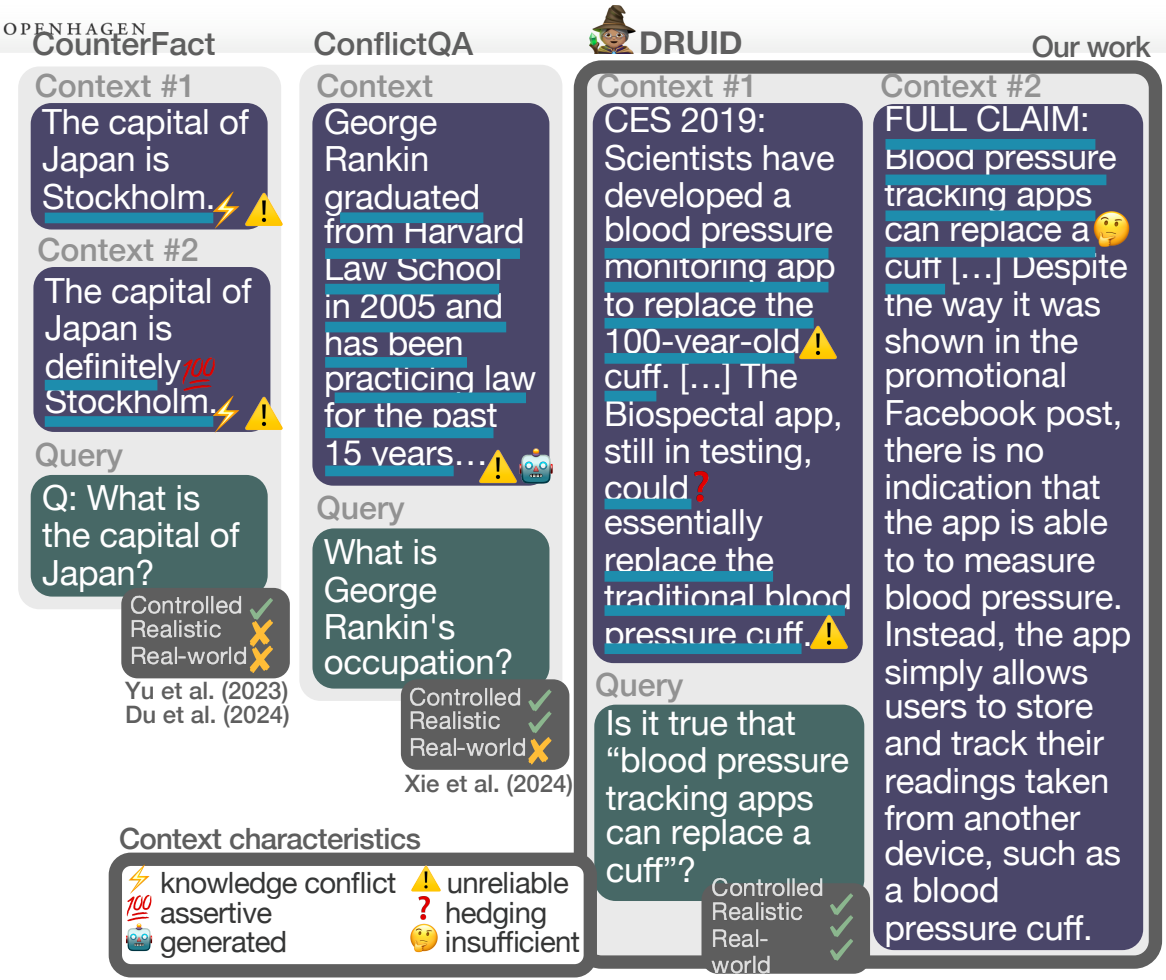
# Context Utilisation of Retrieval-Augmented Generation

- 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 |

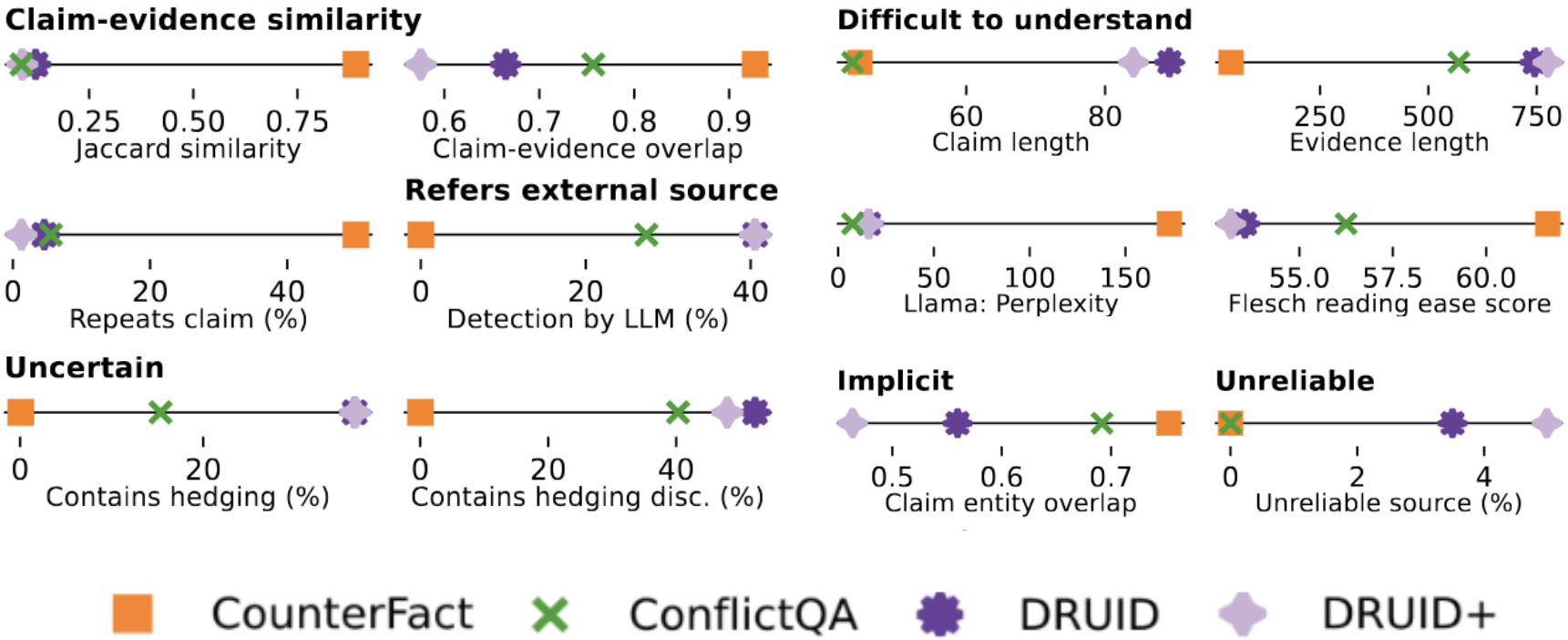
Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

# 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.

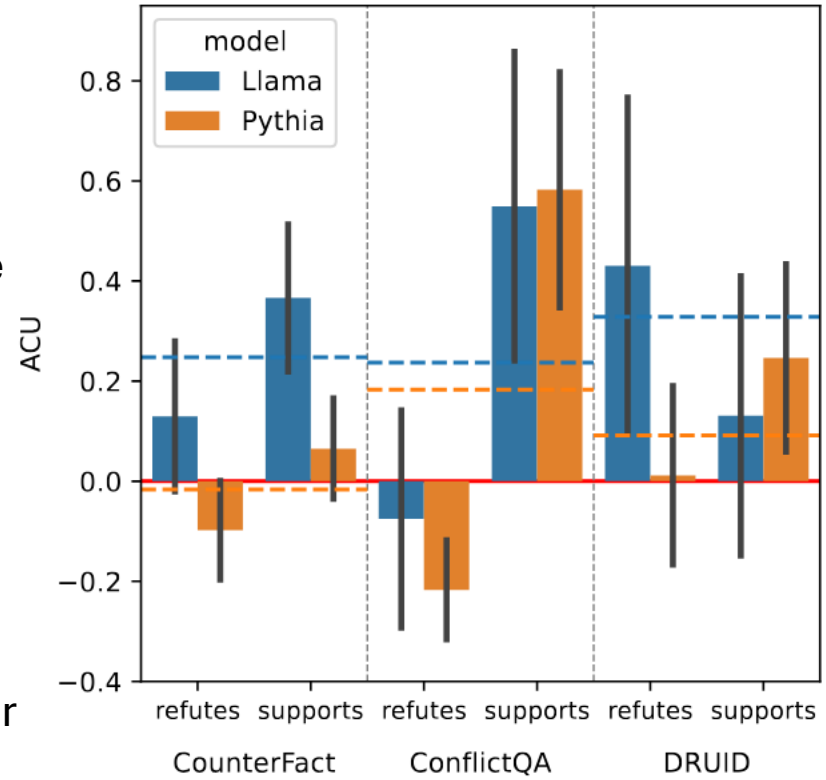
# DRUID content characteristics ctd




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

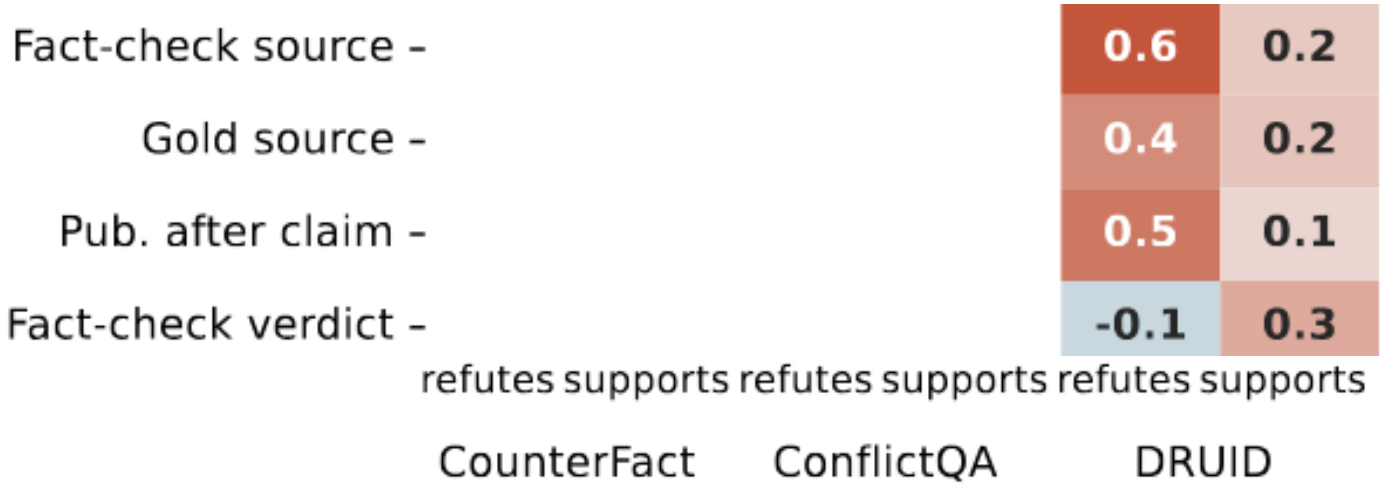
- Context usage (ACU score):
  - Re-scaled difference in salient token probability for different labels for a claim between settings with vs. without evidence
- **Synthetic** datasets:
  - Over-prefer supporting evidence
  - Context repulsion for refuting evidence
  - Generated automatically -> aligned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

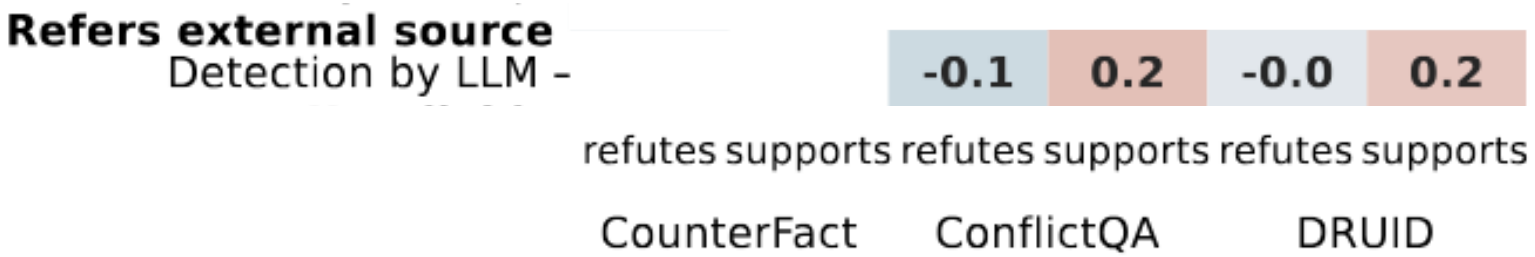
# Influence of content characteristics on RAG

- **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

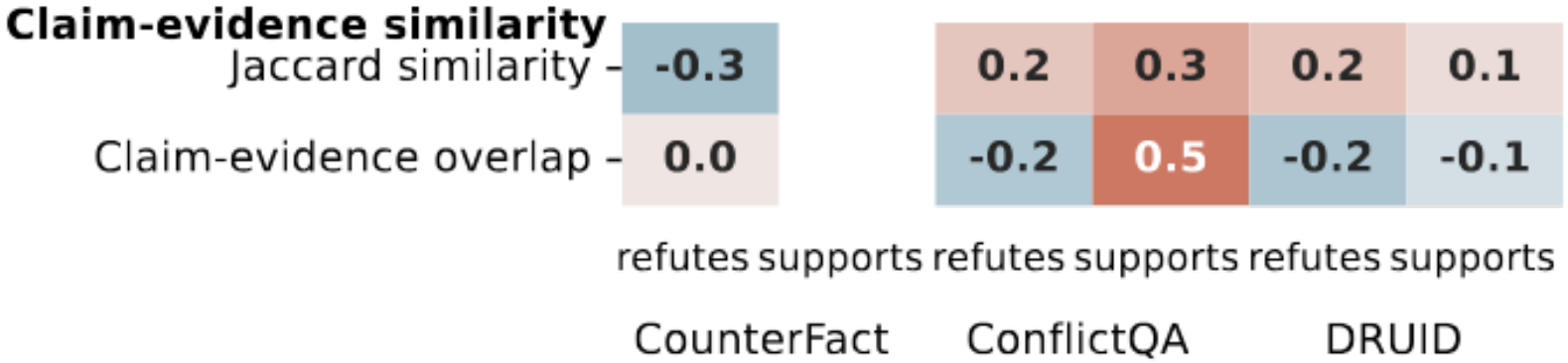
- **References to external sources: low correlations with ACU**
  - Confirms findings of previous work, showing LLM are insensitive to references to external sources




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

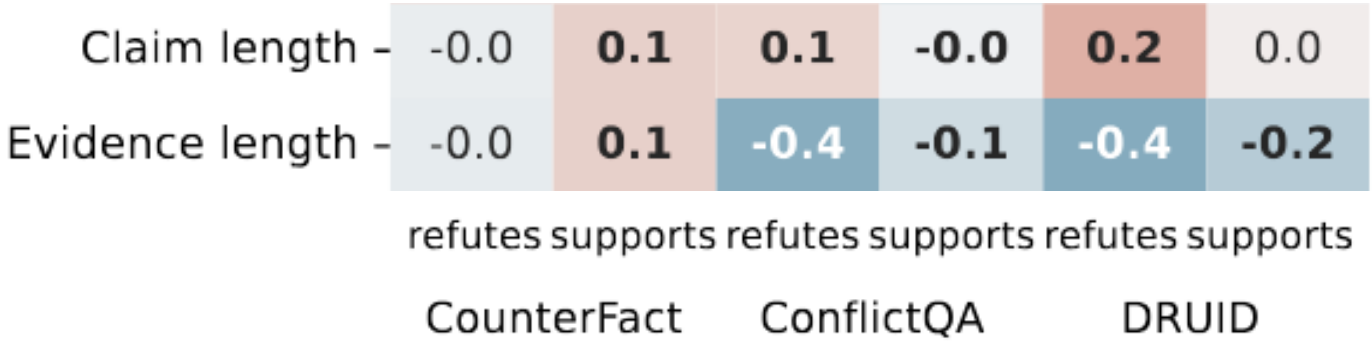
- Correlations with **claim-evidence similarity properties low for DRUID**
  - LLMs prioritise contexts with high query-context similarity -> more difficult in real-world RAG setting




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

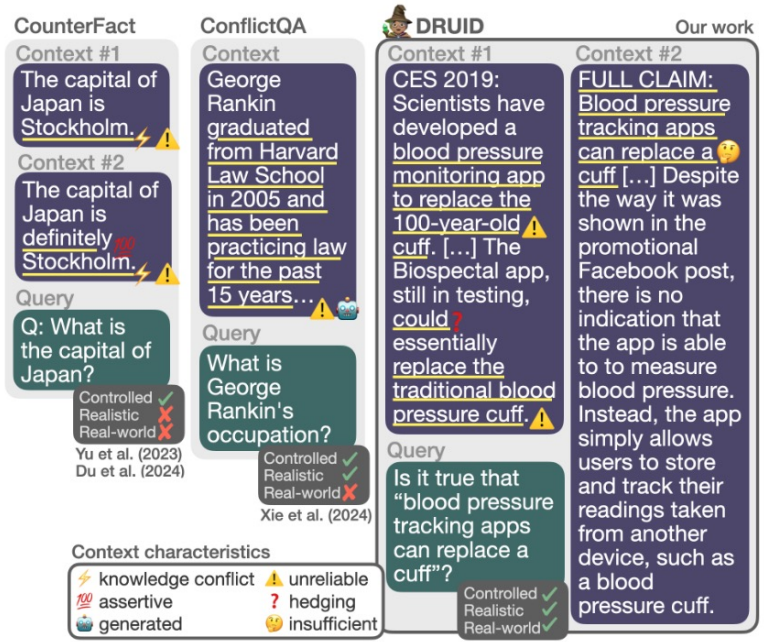
- LLMs **less faithful to long contexts**




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 rarer 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



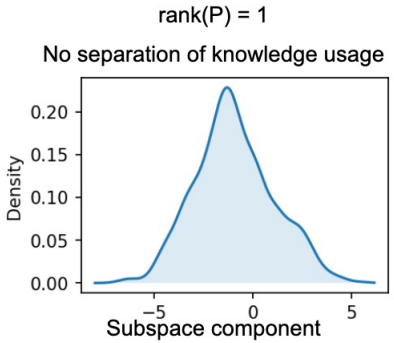

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

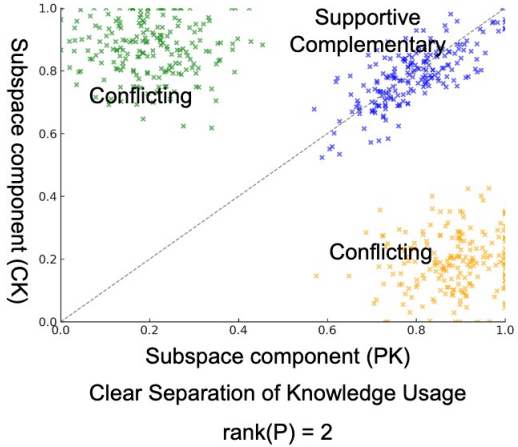
# Multi-Step Knowledge Interaction Analysis

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

**Contributions:**

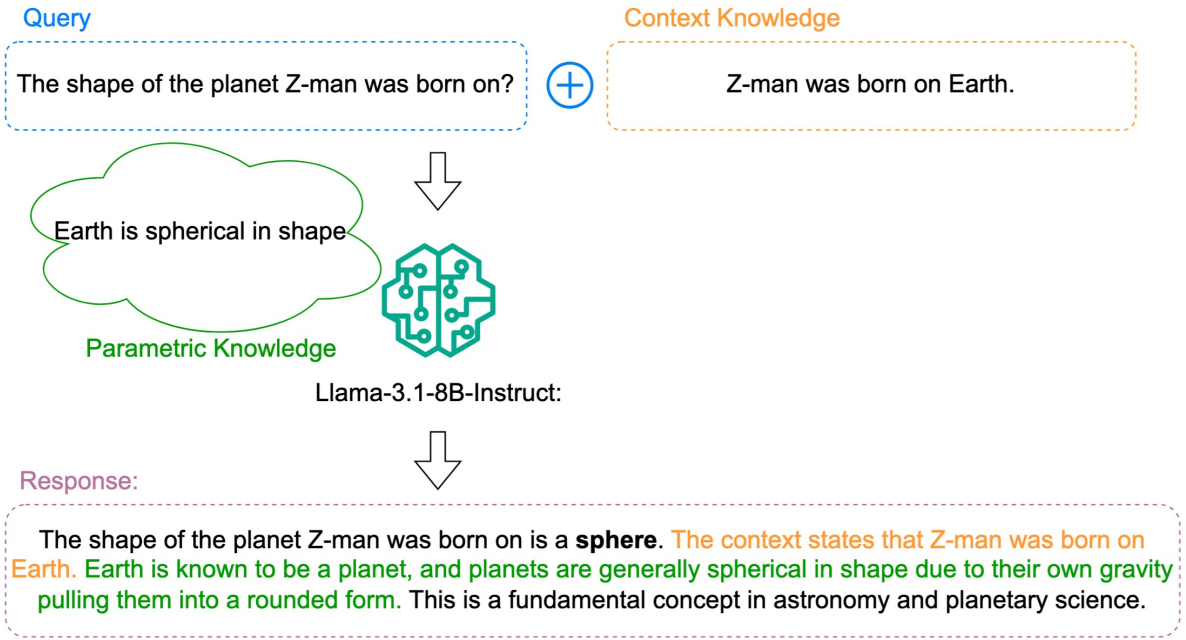
- novel knowledge interaction analysis via rank-2 subspace projection
- application to interaction of long natural language explanation sequences
- novel insights into LLMs' knowledge interaction dynamics




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



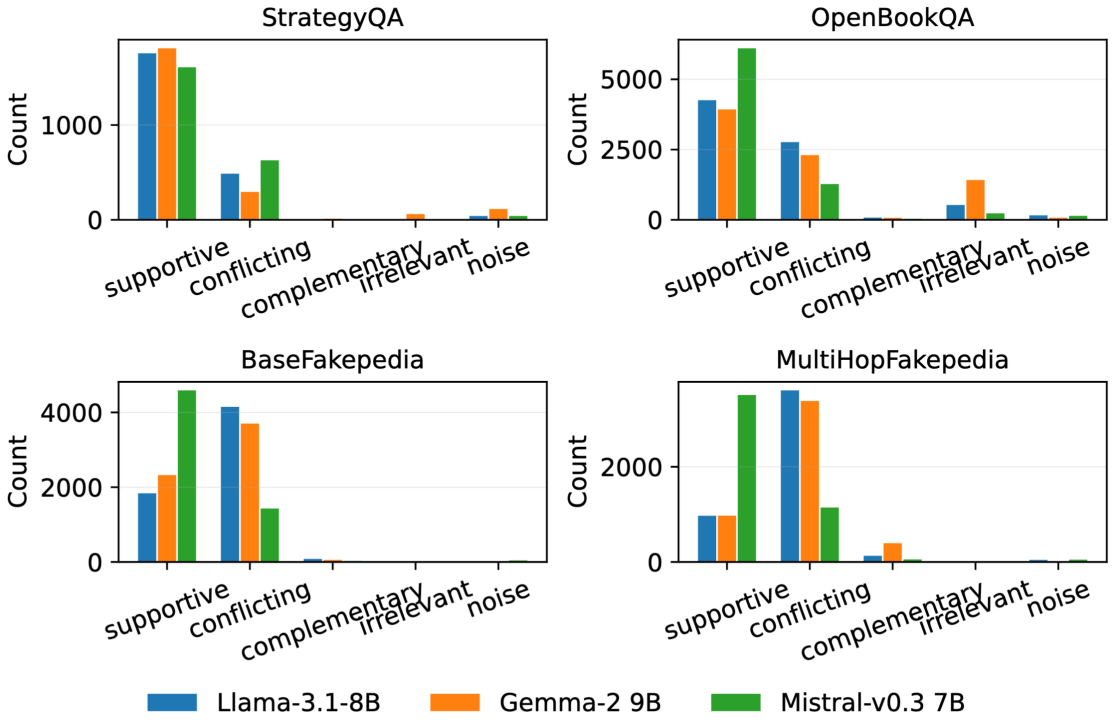

Sekh Mainul Islam, Pepa Atanasova, **Isabelle Augenstein**. Multi-Step Knowledge Interaction Analysis via Rank-2 Subspace Disentanglement. CoRR, abs/2511.01706, November 2025.

# Types of Knowledge Interactions

- **Supportive**: PK and CK reinforce the same outcome, leading to high confidence in the prediction
- **Complementary**: Neither source is sufficient alone, but their combined information yields the final answer
- **Conflicting**: The final answer aligns exclusively with one knowledge source while rejecting the other
- **Irrelevant**: The context contains no semantic information relevant to the query; the model ignores the context entirely and relies solely on PK to formulate a response
- **Knowledge Suppression (noise)**: Both sources agree on an answer, yet the model fails to utilise either, producing an erroneous or unrelated prediction

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

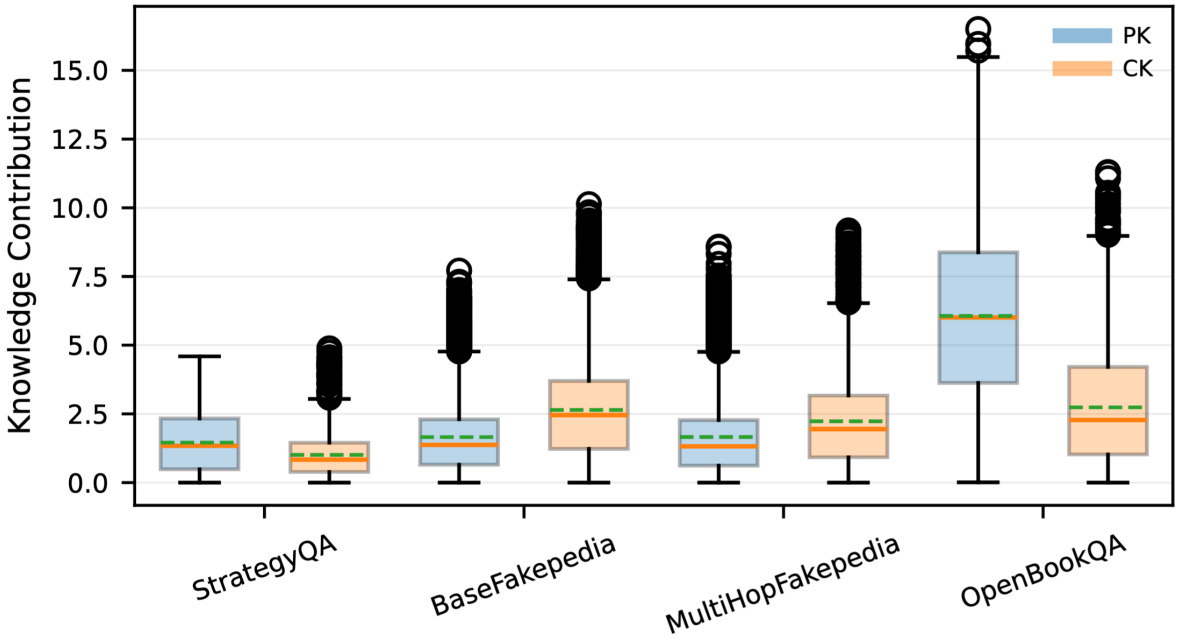
# RQ2: How Do Individual PK and CK Contributions Change Over the NLE Generation for Different Knowledge Interactions?




- 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.

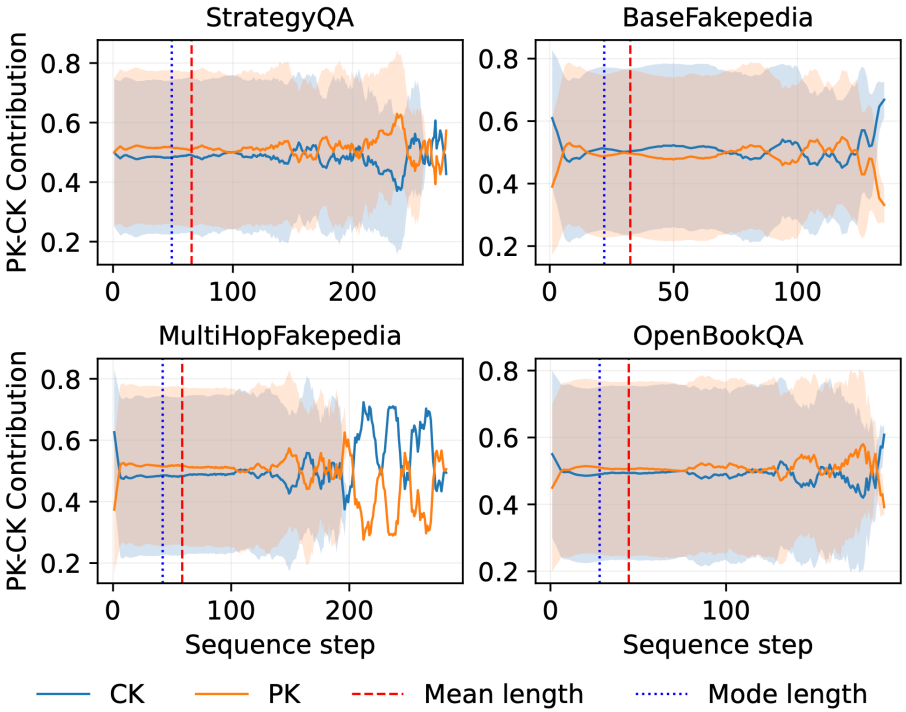
# RQ2: How Do Individual PK and CK Contributions Change Over the NLE Generation for Different Knowledge Interactions?




- 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.

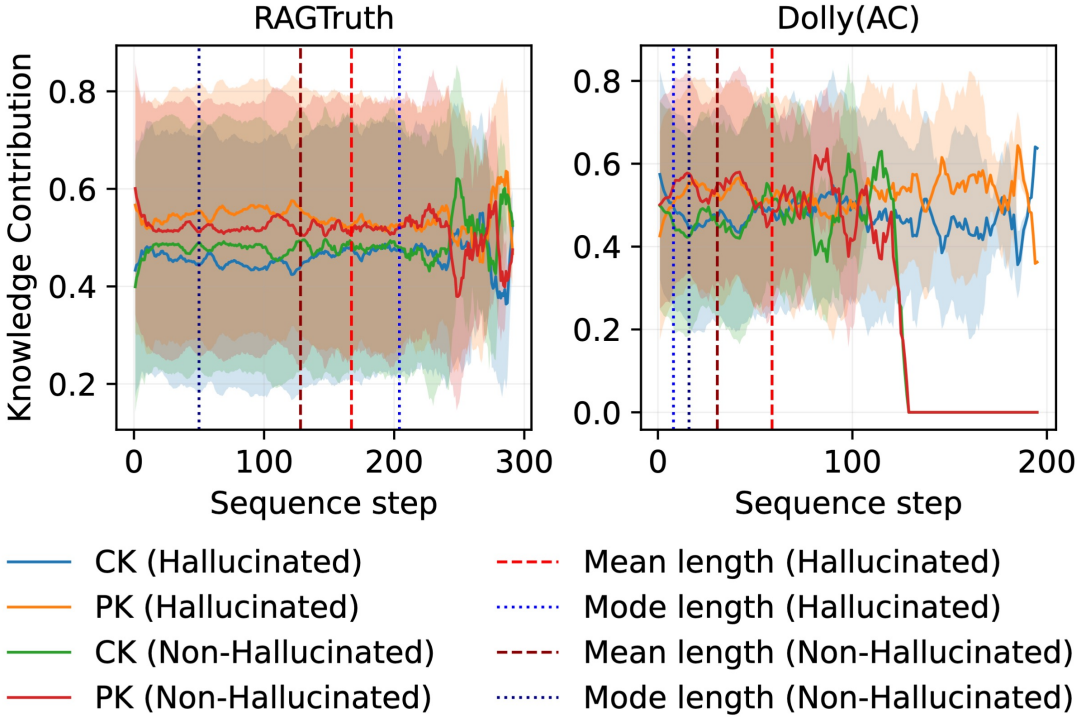
# RQ2: How Do Individual PK and CK Contributions Change Over the NLE Generation for Different Knowledge Interactions?




- 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 behavior

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?

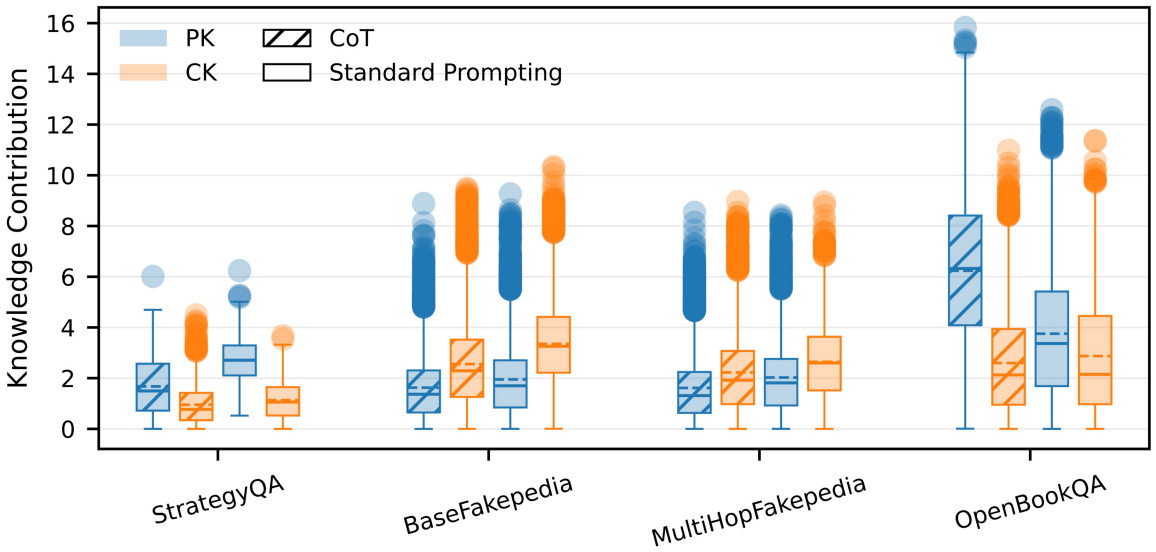



- Gap between PK and CK much higher for hallucinated than for non-hallucinated instances
- Hallucinated answers based more on PK than CK; already visible during early sequence steps

➢ 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. CoRR, abs/2511.01706, November 2025.

# RQ4: How is the CoT Mechanism aligned with the Knowledge Interaction Subspace?




- CoT maintains similar CK alignment compared to standard prompting; reduces PK alignment

➢ CoT operates as a distinct low-rank subspace aligned more with CK

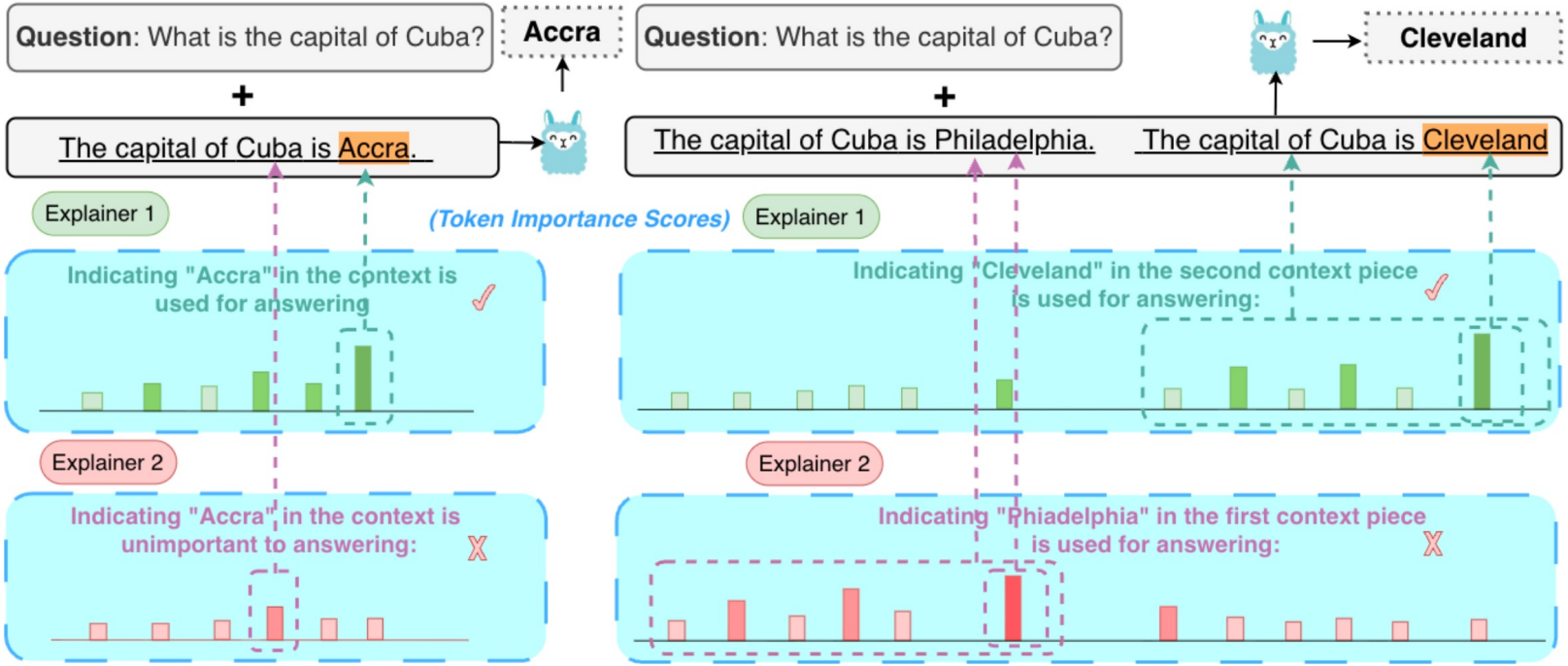
➢ Enhances contextual grounding without fully suppressing PK influence

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
  - Explaining context usage of RAG models
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

# 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, April 2026, to appear.

# Explaining LLMs' Context Usage: Evaluation Setting

- **Conflicting** (single): The context contains an answer that conflicts with PK
- **Irrelevant** (single): The context is irrelevant, but contains a distracting (incorrect) answer token
- **Double-Conflicting** (dual): Two pieces that are conflicting with PK
- **Mixed** (dual): One irrelevant and one conflicting piece

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, April 2026, to appear.

# Explaining LLMs' Context Usage: Evaluation Setting

**Q:** Newport County A.F.C. is headquartered in **MA**: Newport

**Single-Context Setups**

**Input Regime (1) Conflicting C**
Newport County A.F.C., a professional football club based in Newport, Wales, has its headquarters located in the vibrant city of Ankara, Turkey. The club's decision to establish …
**CA:** Ankara

**Input Regime (2) Irrelevant C**
The World Wrestling Entertainment (WWE) is a global entertainment company that is headquartered in Santiago, Chile. Founded in 1952, WWE has become one of the largest …
**CA:** Santiago

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, April 2026, to appear.

# Explaining LLMs' Context Usage: Evaluation Setting

**Q:** Newport County A.F.C. is headquartered in **MA**: Newport

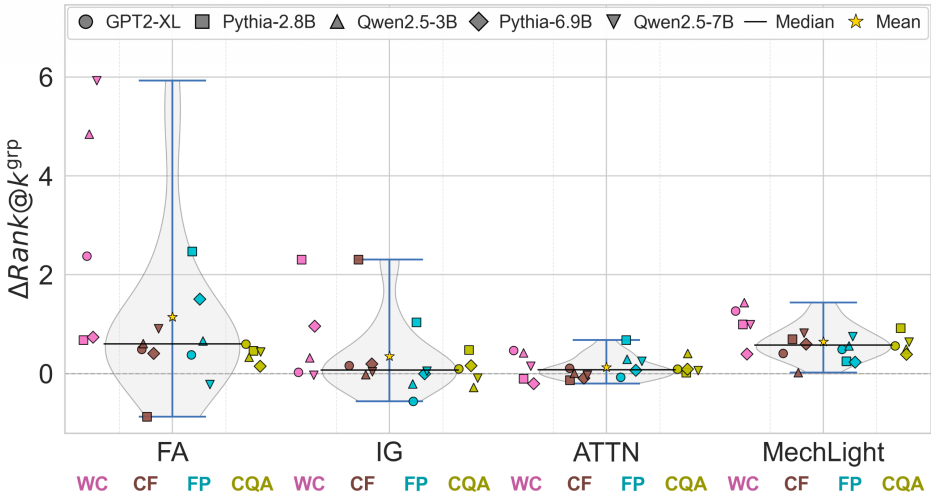
**Dual-Context Setups**

**Input Regime (3) Double Conflict C**
**C P1:** Newport County A.F.C., a professional football club based in Newport, Wales, has its headquarters located in Ankara, Turkey. The club's decision to establish its …
**C P2:** Newport County A.F.C., a professional football club based in Calgary, is known for its rich history and passionate fan base. The club was founded in 1912 and has since become a prominent fixture in the Canadian football scene …
**P1 A:** Ankara **P2 A:** Calgary

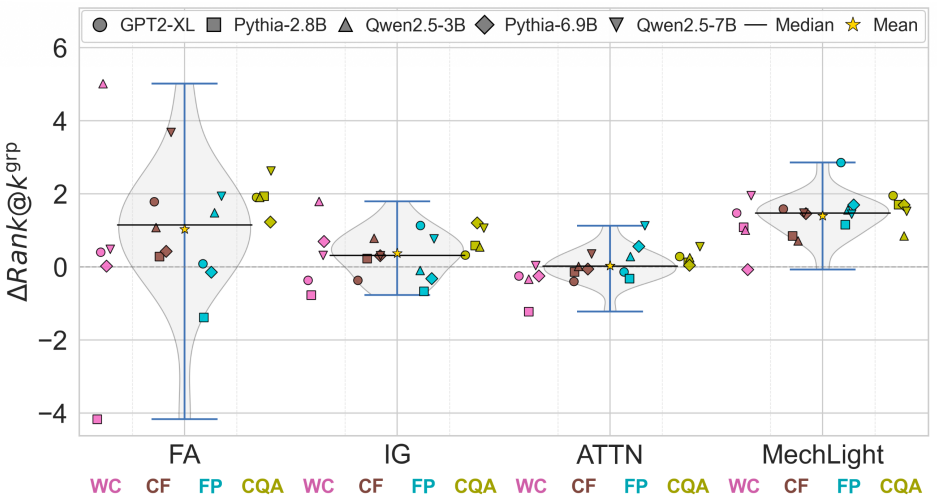
**Input Regime (4) Mixed C** (Irrel. & Conf.)
**C P1:** The World Wrestling Entertainment (WWE) is a global entertainment company that is headquartered in Santiago, Chile. Founded in 1952, WWE has …
**C P2:** Newport County A.F.C., a professional football club based in Newport, Wales, has its headquarters located in Ankara, Turkey. The club's decision to establish its …
**P1 A:** Santiago **P2 A:** Ankara

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, April 2026, to appear.

# RQ1: Does the explanation indicate if the model consulted the context?



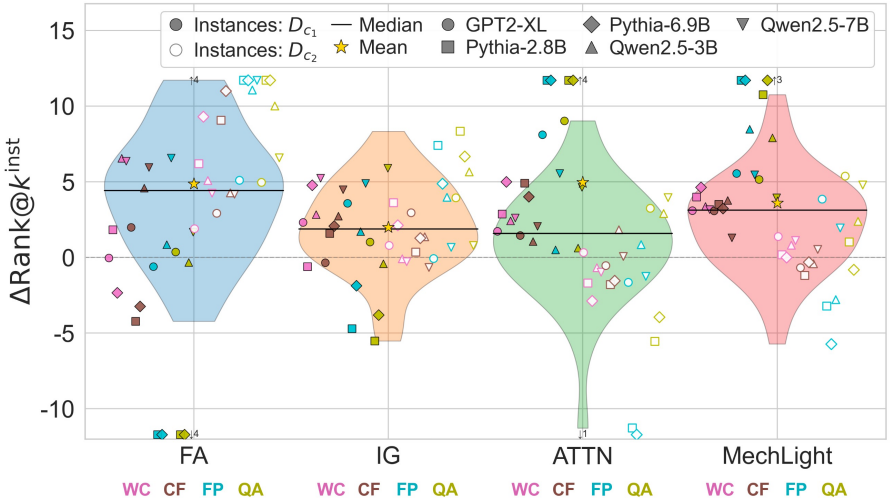

**(a)** Conflicting Context



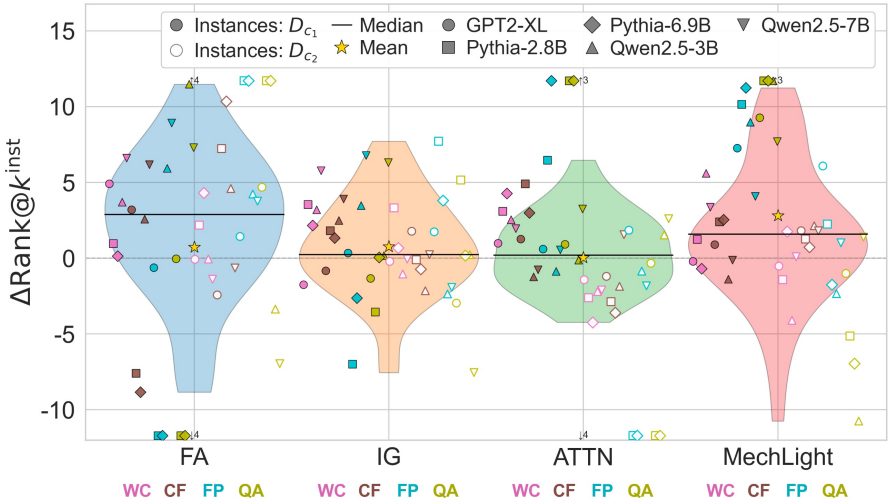

**(b)** Irrelevant Context

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, April 2026, to appear.

# RQ2: Does the explanation show which of two context documents was used?



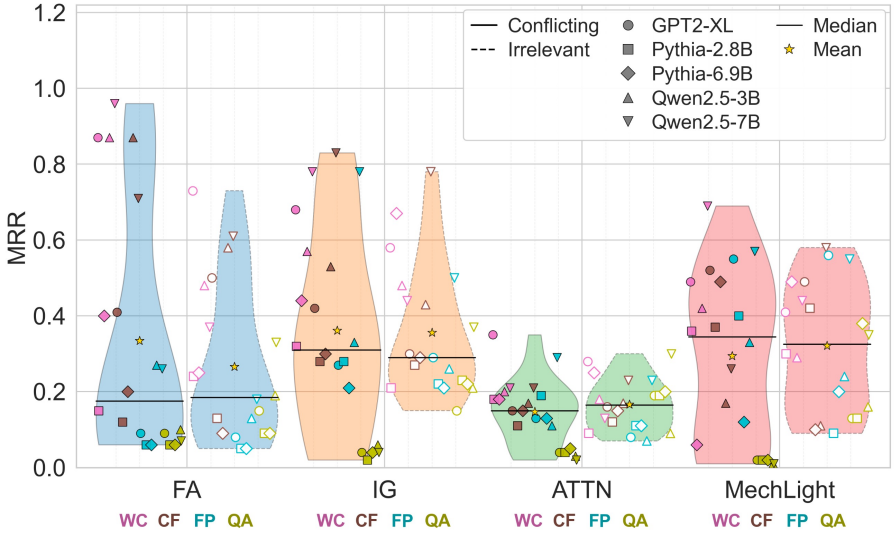

**(a)** Double-Conflicting: Two Conflicting Contexts



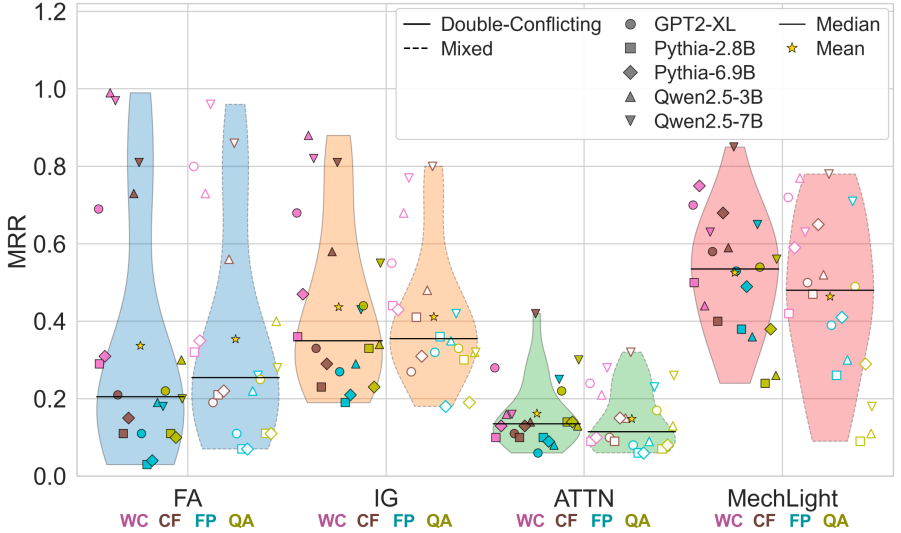

**(b)** Mixed: One Irrelevant and One Conflicting Context

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, April 2026, to appear.

# RQ3: Does the explanation pinpoint the exact context part(s) employed for the generated answer?




**(a)** Conflicting & Irrelevant




**(b)** Double-Conflicting & Mixed

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, April 2026, to appear.

# Explaining LLMs' Context Usage: Error Analysis

| Dataset | Prompt slice | Model answer |
|---|---|---|
| **Obs. 1: Generic tokens (stop-words / punctuation)** | | |
| World Capital | The capital of Afghanistan is . Valletta … Q: What is the capital of Afghanistan? A: | Valletta |
| World Capital | The capital of Algeria is . Sukhumi … Q: What is the capital of Algeria? A: | Sukhumi |
| CounterFact | Fact : Accra, the capital city of Lebanon. Q: Accra, the capital city of A: | Lebanon |
| **Obs. 2: Nearby descriptors/entity names** | | |
| ConflictQA | Trade paper profile begins Veteran producer and studio head Dore Schary … Q: Who directed Act One? | Dore Schary |
| ConflictQA | Effects journal states Visual effects expert Bruno was hired for "Virus" … Q: Who was the director of Virus? | John Bruno |
| Fakepedia | Apple Pay white-paper: the ground breaking payment service launched with Intel hardware … Q: Apple Pay, a product created by | Intel |
| **Obs. 3: Question focus** | | |
| World Capital | … Q: What is the capital of Albania? A: | Berlin |
| ConflictQA | … Q: Who was the director of "Virus"? A: | John Bruno |
| CounterFact | … Q: What is the capital of Burgundy? A: | Bangkok |

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, April 2026, to appear.

# Explaining LLMs' Context Usage: Findings

- Across five LMs and four commonly used context-usage datasets, we find:
  - MechLight (attention Head Scoring + highlight explanation extraction) performs best across all context scenarios
  - Systematic limitations across all highlight explanations:
    - (i) length sensitivity – HE accuracy degrades as context grows
    - (ii) position biases under dual-context inputs: FA/IG tend to favour later (near-question) pieces, while ATTN/MechLight favour earlier pieces

- Surprisingly, the widely used IG and ATTN overall exhibit poor accuracy
  - ➢ useless in revealing the model's context utilisation

➢ Urgent need for explanation techniques that maintain accuracy at scale and overcome positional biases in 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, April 2026, to appear.

# 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
  - Explaining context usage of RAG models
  - Context manipulation techniques
- **Conclusion**
  - Wrap-up and outlook

# Benchmarking context usage manipulation techniques

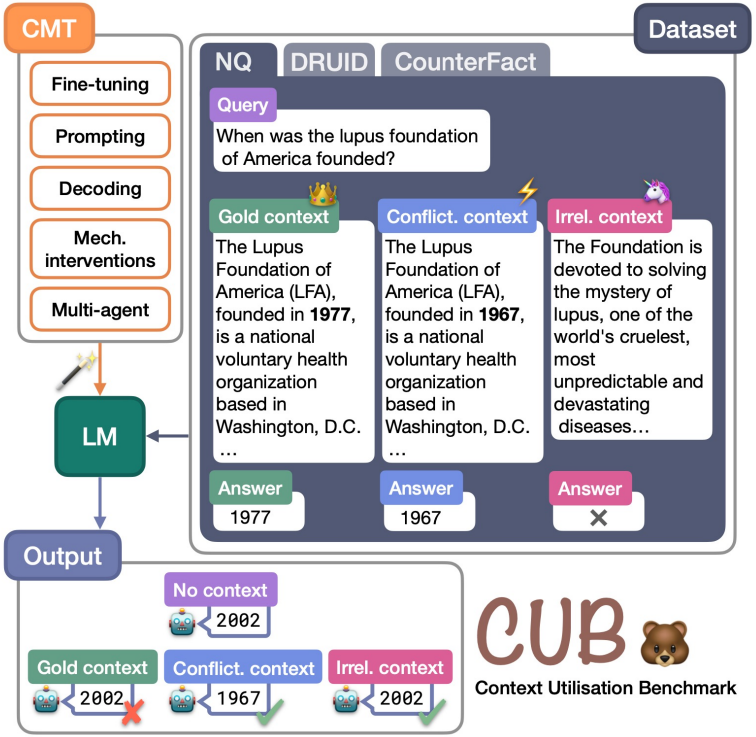
- Previous context usage experiments show that LLMs:
  - Struggle with more complex and long contexts
  - Can easily be distracted by irrelevant contexts due to context-memory conflicts

- 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.

# Context usage manipulation via fine-tuning

- Idea:
  - Update model parameters to modify context utilisation
  - Fine-tuning on distracting contexts can improve robustness to distracting contexts
- Approach:
  - Obtain parametric answers by querying without contexts
  - Select the questions that the LM answered correctly and pair them with irrelevant and empty contexts
- The fine-tuning data thus contains contexts that can be irrelevant, counterfactual or empty
- Fine-tune LM to generate answers aligned with the provided context
- Thus, for irrelevant context, the LM learns to ignore the context and output its parametric answer

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.

# Context usage manipulation via prompting

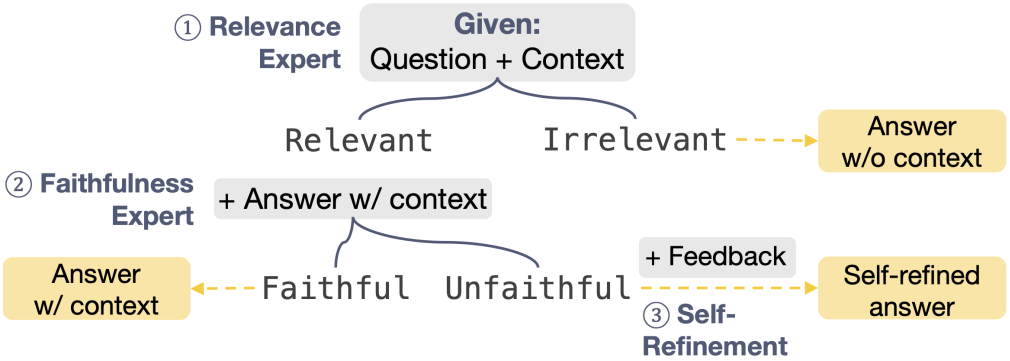
- Prompt tuning for different datasets, e.g.

> “Answer the following questions based on the context below.
> Question: […]
> Context: […]
> Answer:
> ”

> “Answer the question. Only answer with the answer. Examples of questions and desired answers are given below.
> […]
> Now, answer the following question (only with the answer):
> …
> “

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.

# Context usage manipulation via multi-agent refinement




**Algorithm 1** Multi-agent

**Given:** question $q$, context $c$
**Stage1: Relevance Assessment**
Predict $f_{\text{rel}} \sim \text{LM}_{\text{rel}}(f_{\text{rel}} \mid q, c)$
**if** $f_{\text{rel}} = \texttt{Relevant}$ **then**
    Proceed to Stage 2
**else**
    **return** $\text{LM}(a \mid q)$ ▷ Answer w/o $c$
**end if**
**Stage 2: Context-Faithfulness**
Predict $a_c \sim \text{LM}(a_c \mid q, c)$
Predict $f_{\text{faith}} \sim \text{LM}_{\text{faith}}(f_{\text{faith}} \mid q, c, a_c)$
**if** $f_{\text{faith}} = \texttt{Faithful}$ **then**
    **return** $a_c$ ▷ Answer w/ $c$
**else**
    Proceed to Stage 3
**end if**
**Stage 3: Self-Refinement**
**return** $\text{LM}(a \mid q, c, a_c, f_{\text{faith}})$ ▷ Self-Refined

# Context usage manipulation via multi-agent refinement

Relevance Assessment (NQ)

```
You are a relevance assessment expert. Your task
     is to evaluate whether the provided context
     is relevant to the question.

Context: {context}
Question: {question}

If the provided context is relevant to the
    question, answer "Relevant", otherwise
    answer "Irrelevant". Do not rely on your own
     knowledge or judge the factual accuracy of
    the context.
Answer:
```

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.

# Context usage manipulation via multi-agent refinement

Context faithfulness (CounterFact and NQ)

```
You are a context-faithfulness expert. Your task
     is to evaluate whether the proposed answer
    faithfully uses the information in the
    provided context.

Context: {context}
Question: {question}
Proposed answer: {response}

Does the answer faithfully reflect the content
    of the context? Do not rely on your own
    knowledge or judge the factual accuracy of
    the context. Please explain briefly.

Feedback:
```

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.

# Context usage manipulation via multi-agent refinement

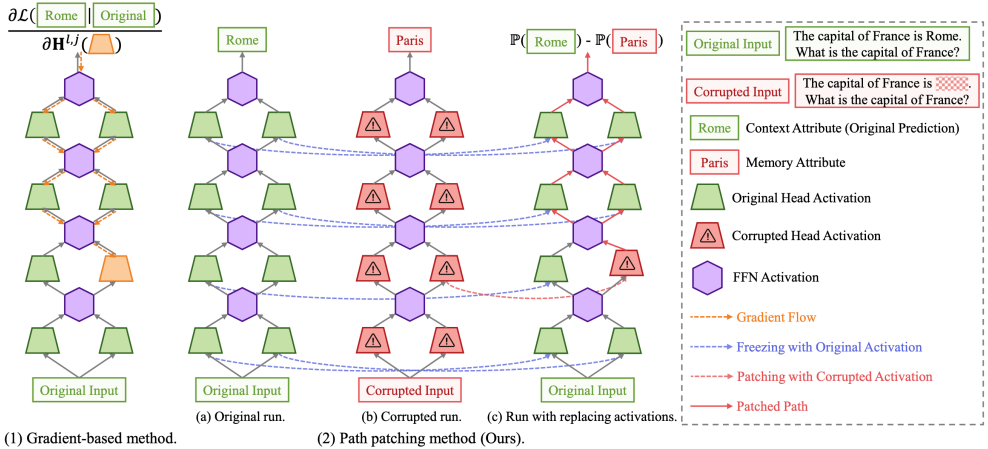
Self-refinement (NQ)

```
Your task is to generate the best possible final
     answer to the question, based on the expert
     feedback.
You may keep the original proposed answer if it
    is correct, or revise it if the feedback
    suggests it is incorrect or unsupported.
Generate only the final answer. Do not include
    any explanation or repeat the prompt.

{Two demonstrations}

Context: {context}
Question: {question}
Proposed answer: {response}
Feedback on context faithfulness: {feedback}
Final answer:
```

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.

# Context usage manipulation via mechanistic interventions (PH3)

1) Identification of attention heads responsible for context or memory reliance via path patching
2) Pruning the identified attention heads for increased memory or context usage




Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models. Proceedings of ACL 2024.

# Context usage manipulation via context-aware contrastive decoding (ACD, COIECD)

- **Adjust model's output distribution** based on distribution for answer to query with vs. without context

- Adaptive-contrastive decoding (ACD): use entropy-based weighting to adaptively ensemble parametric and contextual distributions, addressing stability to irrelevant context
- Contextual information-entropy constraint decoding (COIECD): improve faithfulness to conflicting context without compromising performance when no conflict exists

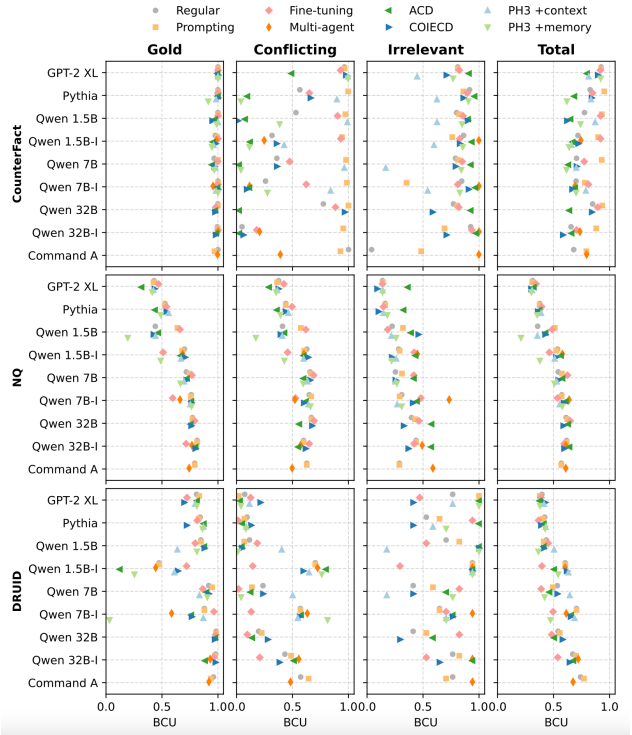
Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.

# 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?



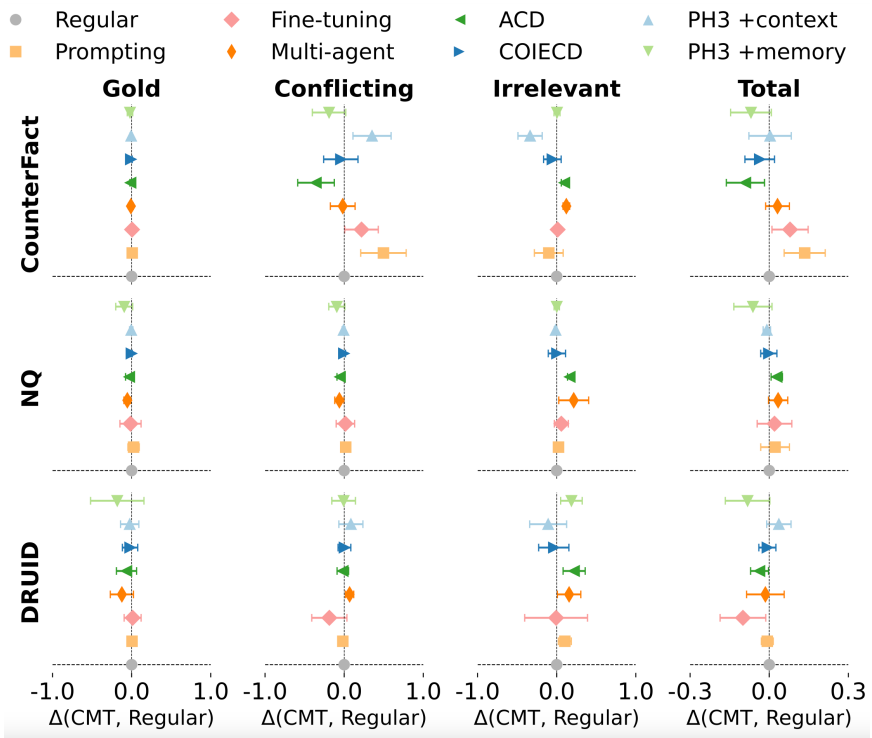

**Binary context utilisation (BCU) score**:

- 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
  - Explaining context usage of RAG models
  - 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 easily persuaded 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

- Can highlight **explanations show which context (parts) were utilised**?
  - Highlight explanations can overall indicate if model consulted the context (though large variability)
  - Can overall not indicate which of several contexts was consulted
  - Can overall indicate which context parts were used
  - Limitations: large variability across LLMs and datasets, degredation of performance for long contexts, position bias of explanation methods

- 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

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, April 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. 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.

# EACL 2027

Isabelle Augenstein
University of Copenhagen

**General Chair**

Malvina Nissim
University of Groningen

**Program Chair**

Roi Reichart
Israel Institute of Technology

**Program Chair**

Sara Tonelli
Fondazione Bruno Kessler

**Program Chair**

# EACL 2027

Athens, Greece • Αθήνα, Ελλάδα
March, 9–14, Μαρτίου

# 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.

## Yoonna Jang

**Postdoc**

Yoonna's research interests in language generation, factualit 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 NLP and Computational Social Science.

## Sara Vera Marjanovic

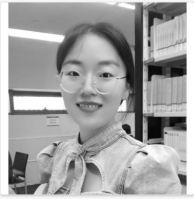
**PhD Student**

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 interest include enhancing explainability fact-checking and transparency 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 misleading text. She is a PhD student at Aarhus University and co-advised by Isabelle

## 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.

## Jean Seo

**PhD Student**

Jean's research interests include improving the safety of language models through explainability and evaluation.

+ 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 the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

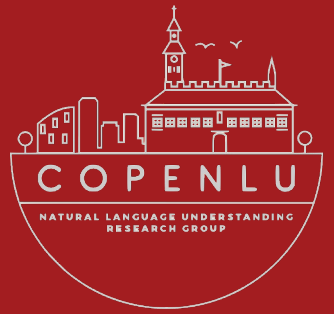
Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of ACL 2025, July 2025.

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

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, April 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. CoRR, abs/2505.16518, May 2025.

# Thank you for your attention! Questions?




We're hiring!