Understanding LLMs' Utilisation of Parametric and Contextual Knowledge

Isabelle Augenstein

ECIR 2025 - Karen Spärck Jones Award lecture
8 April 2025

COPENLU

Funded by the European Union

erc
European Research Council
Established by the European Commission

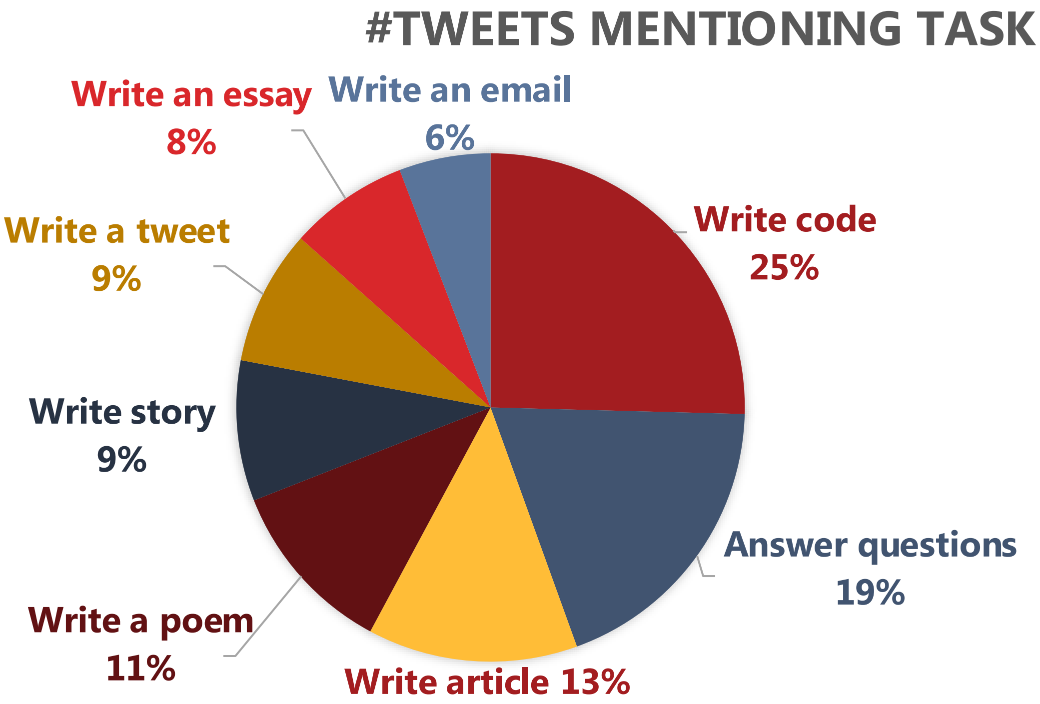
bcs
The Chartered Institute for IT

Bloomberg

UNIVERSITY OF COPENHAGEN

LLM usage is ubiquitous

Website	Total visits
Amazon	3.1 billion
WhatsApp	3.8 billion
X	4.8 billion
ChatGPT	**5.2 billion**
Wikipedia	7 billion
…	…
Google	139.9 billion

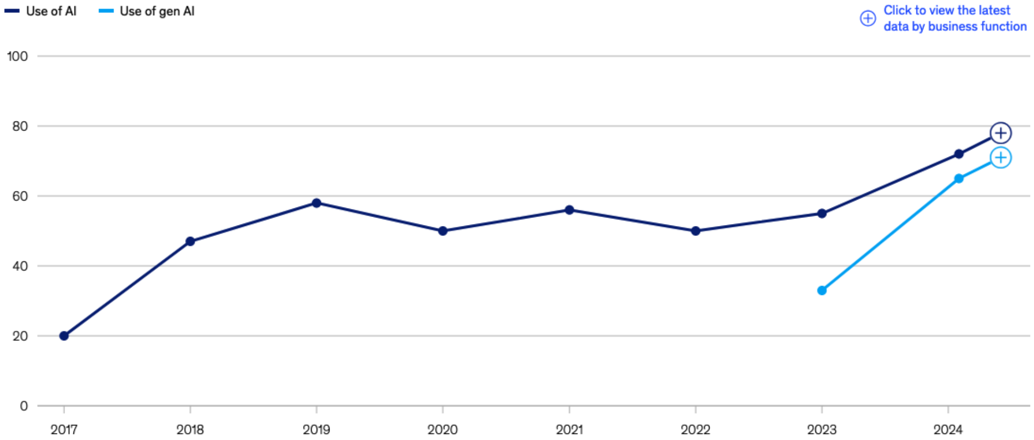


Exploding Topics. Number of ChatGPT Users (2025), 25 March 2025. https://explodingtopics.com/blog/chatgpt-users
Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, Gualtiero Fantoni. Future applications of generative large language models: A data-driven case study on ChatGPT. Technovation Volume 133, May 2024, 103002.

It is transforming organisations

Organizations' use of AI has accelerated markedly in the past year, after years of little meaningful change.

Organizations that use AI in at least 1 business function,[1] % of respondents



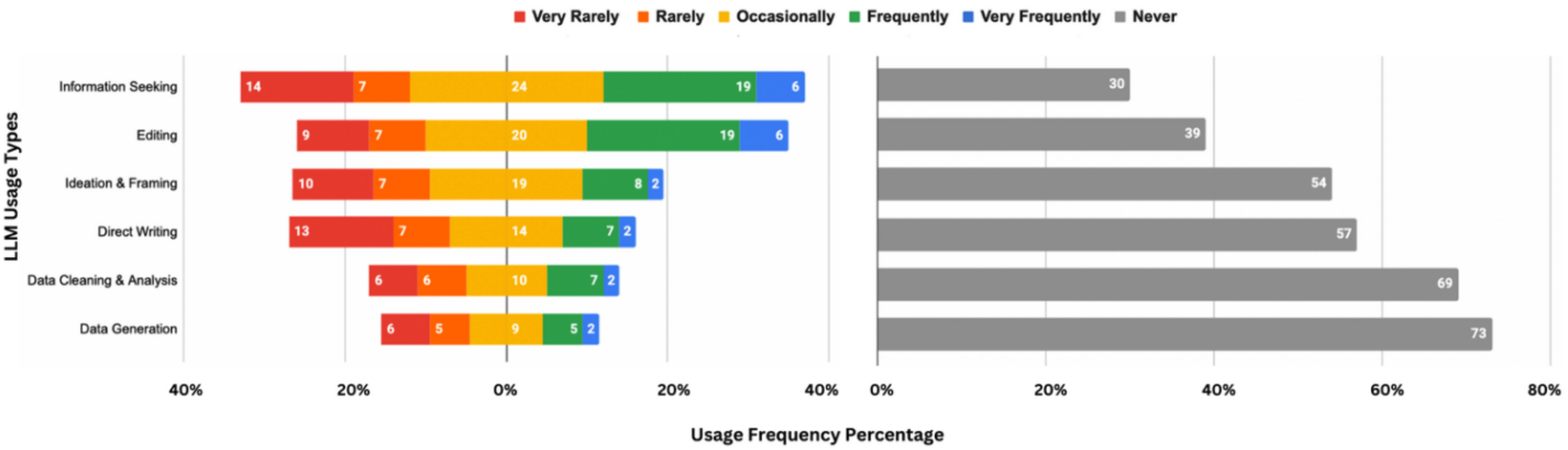
Organizations across industries have begun to use gen AI in marketing and sales, though other uses vary by industry.

Business functions in which respondents' organizations are regularly using gen AI, by industry,[1] % of respondents

	Technology	Professional services	Advanced industries	Media and telecom	Consumer goods and retail	Financial services	Healthcare, pharma, and medical products	Energy and materials	Overall
Marketing and sales	55	49	48	45	46	40	29	33	42
Product and/or service development	39	41	39	26	21	25	22	17	28
IT	31	16	26	22	20	24	30	26	23
Service operations	30	23	24	37	13	26	14	13	22
Knowledge management	26	34	17	26	12	16	24	13	21
Software engineering	36	9	17	30	8	20	13	8	18
Human resources	16	17	13	22	8	11	7	16	13
Risk, legal, and compliance	12	9	6	6	11	21	5	9	11
Strategy and corporate finance	14	14	21	10	7	7	6	5	11
Supply chain/ inventory management	10	4	15	3	14	4	2	6	7
Manufacturing	5	3	13	3	8	0	5	7	5
Using gen AI in at least 1 function	88	80	79	79	68	65	63	59	71

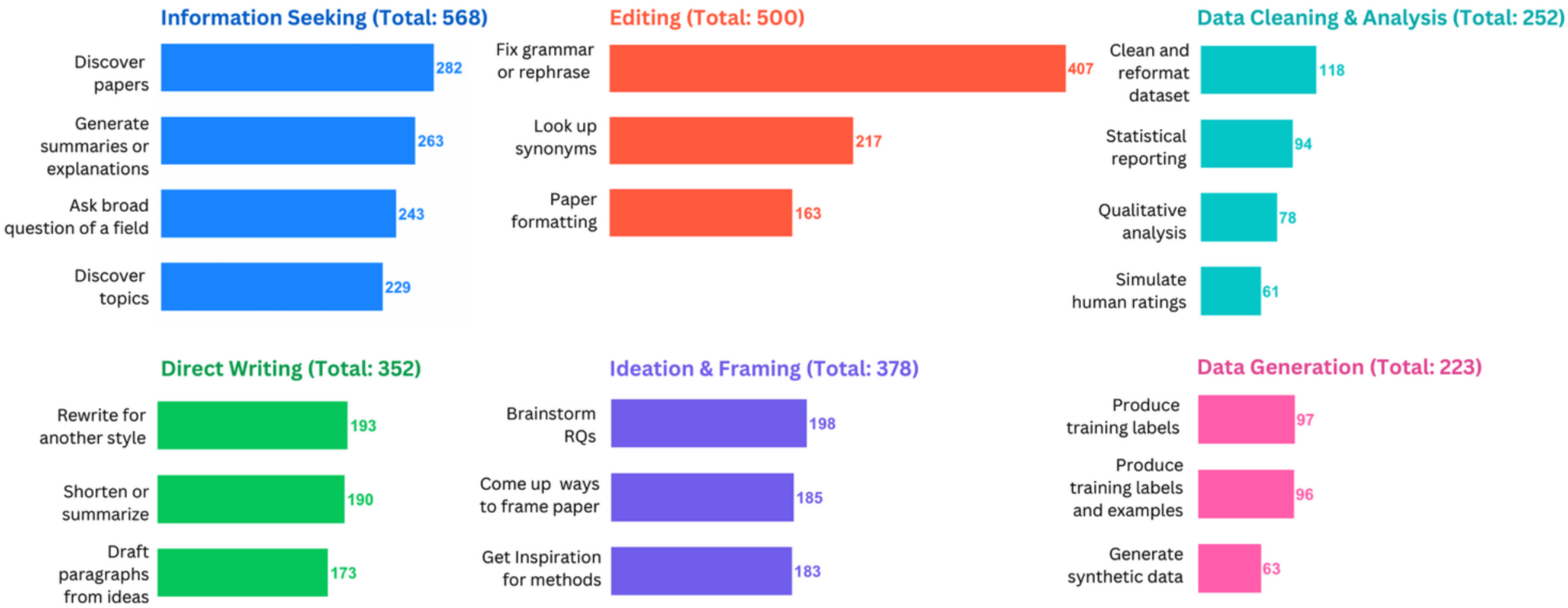
McKinksey. The state of AI: How organizations are rewiring to capture value. 12 March 2025.

And research itself



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

And research itself



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

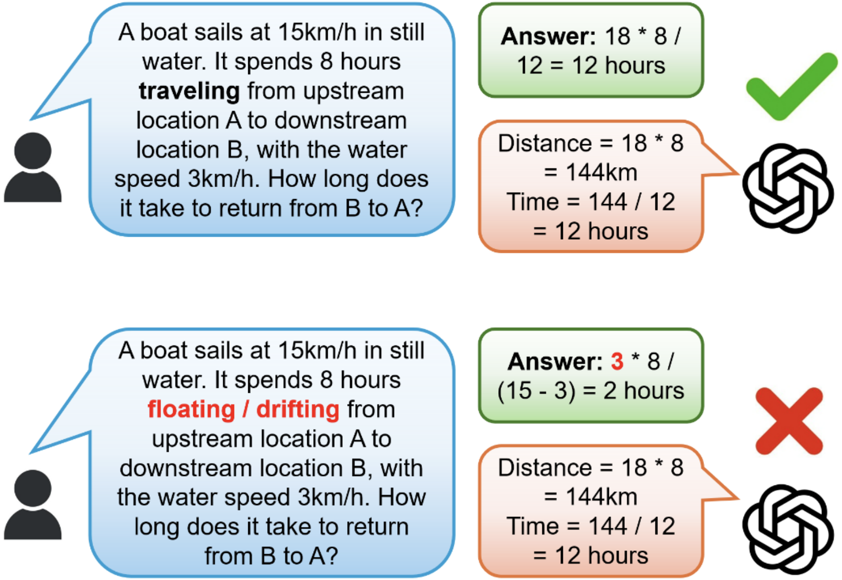
Are we seeing the emergence of AGI?

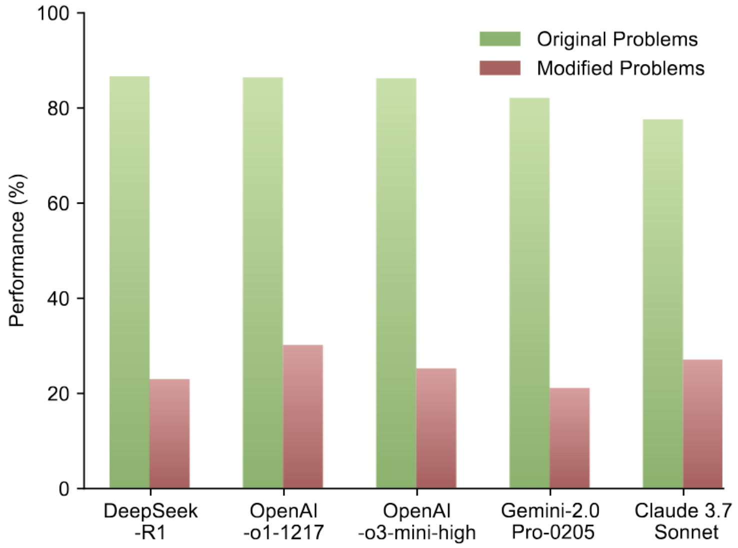
NO

Are we seeing the emergence of AGI?

- LLMs show high performance generally, but display several fundamental shortcomings

- Outperform previous models on various NLP tasks on existing benchmarks
  - ⚠ : high **dataset contamination** -> most test sets seen at training time
  - Drastic performance drops when performing small alterations to wording





Are we seeing the emergence of AGI?

- LLMs show high performance generally, but display several fundamental shortcomings

- Outperform previous models on various NLP tasks on existing benchmarks
  - ⚠ : high **dataset contamination** -> most test sets seen at training time
  - Drastic performance drops when performing small alterations to wording

- Poor performance on low- and very low-resource languages
- Poor at most types of reasoning

- **Many factual errors** due to lack of access to an external knowledge base

- Take-aways:
  - LLMs are excellent at recitation, not at reasoning
  - LLMs are multi-task learners, but not AGI models

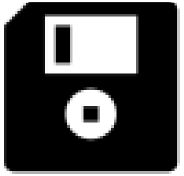
Bang et al. (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In ICJNLP/AAACL 2023.
Yan et al. (2025). Recitation over Reasoning: How Cutting-Edge Language Models Can Fail on Elementary School-Level Reasoning Problems? Arxiv, abs/2504.00509, April 2025.

Factuality Challenges of Large Language Models


Citation Gaps


Truthfulness


Fluent Style


Outdated Knowledge


Grounding Deficiency

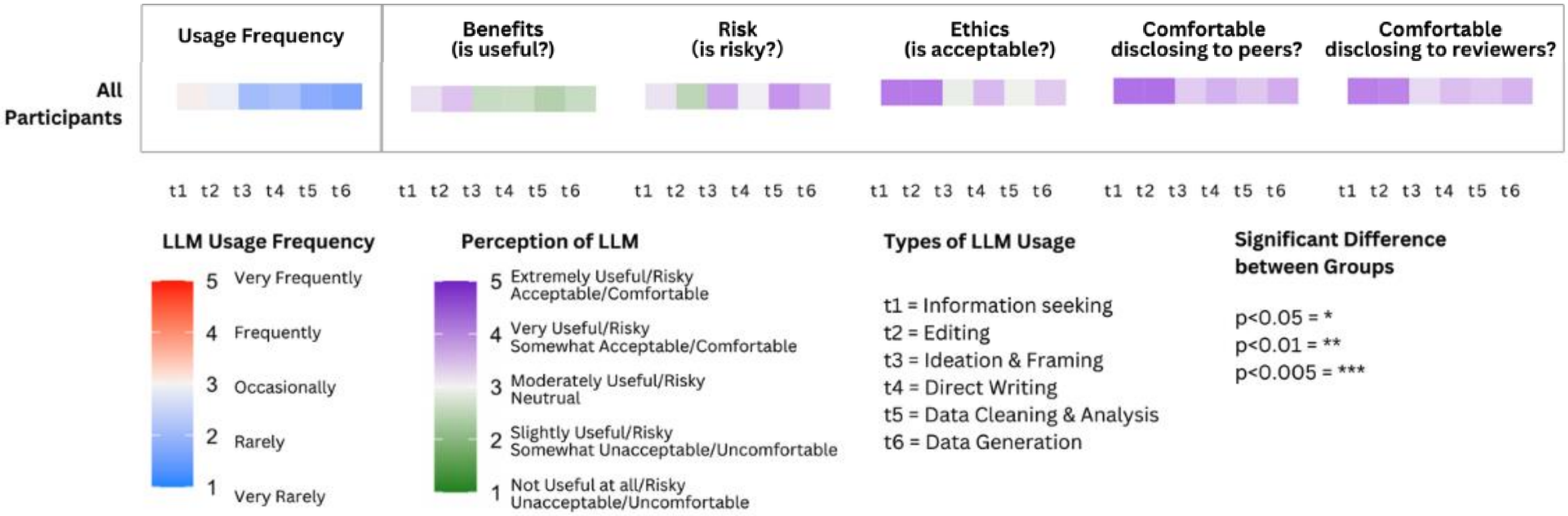

Confident Tone


Halo Effect


Unreliable Evaluation

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

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](). ArXiv, abs/2411.05025.

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."* *"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."* *"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."* *"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. ArXiv, abs/2411.05025.

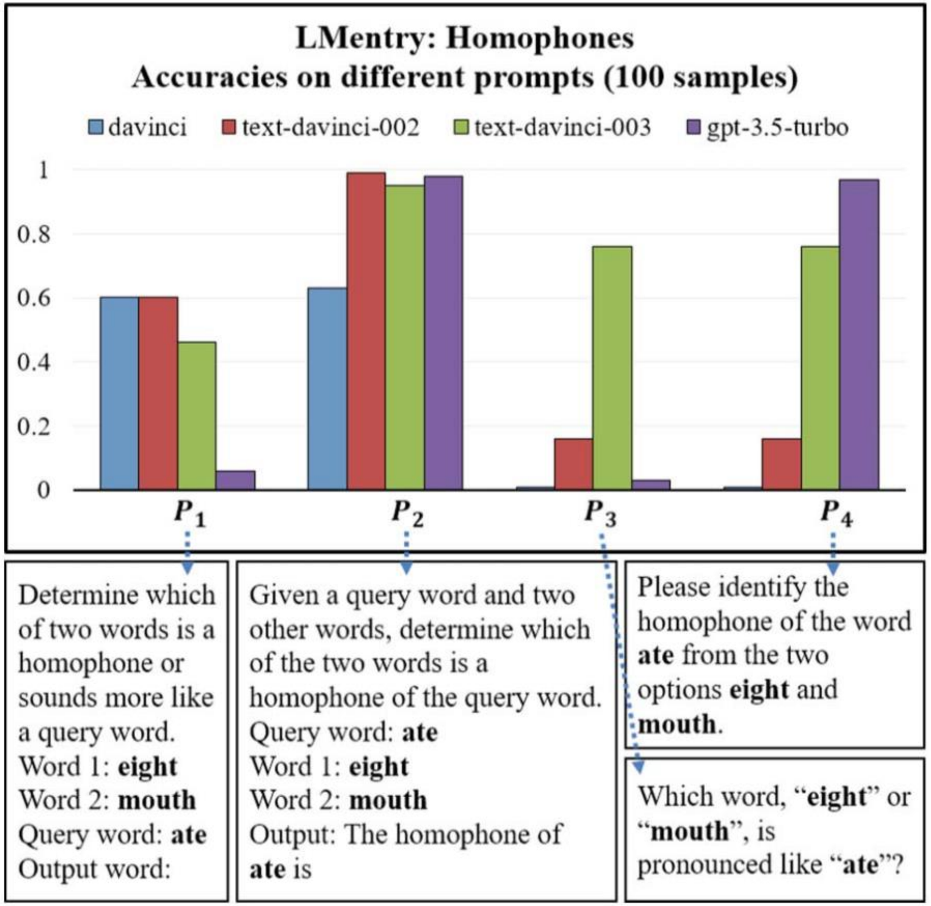
How to address factuality issues of LLMs?

Improving consistency

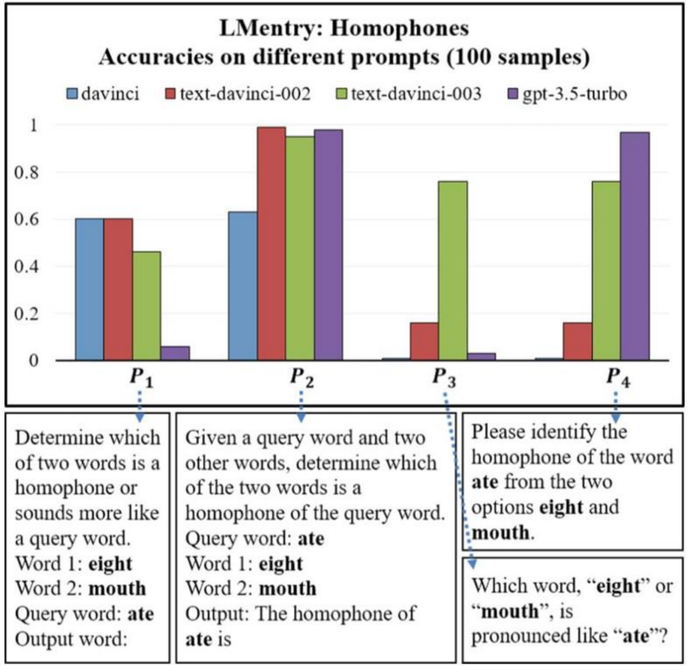
- Self-consistency checking
- Chain-of-thought prompting
- Continual learning
- Knowledge editing

Problems:

- Knowledge editing is difficult
  - -- Ripple effects of knowledge editing
  - -- How to even know what knowledge to edit?
  - -- Risk of removing long-tail knowledge
- LLMs are not very self-consistent
  - -- Prompt instability
  - -- No single ”personality” or ”right answer”



How to address factuality issues of LLMs?



Improving consistency

- Self-consistency checking
- Chain-of-thought prompting
- Continual learning
- Knowledge editing

Problems:

- Knowledge editing is difficult
  - -- Ripple effects of knowledge editing
  - -- How to even know what knowledge to edit?
  - -- Risk of removing long-tail knowledge
- LLMs are not very self-consistent
  - -- Prompt instability
  - -- No single "personality" or "right answer"

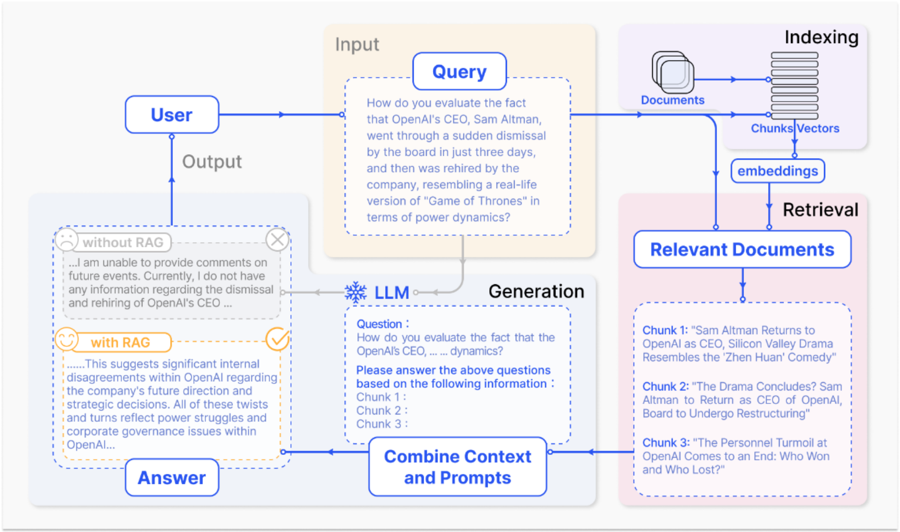
➢ *LLMs are used for both creative and information-seeking tasks*
➢ *Knowledge-intensive tasks are highly context-dependent*
➢ *Internal consistency checking only partly address issues for information-seeking tasks*

Augenstein et al. (2024). Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.
Mizrahi et al. (2024). State of What Art? A Call for Multi-Prompt LLM Evaluation. In TACL.

How to address factuality issues of LLMs?

Combination with external knowledge

- Detecting and correcting factual mistakes at inference time
- Modularised knowledge-grounded framework
- Retrieval-augmented generation (RAG)

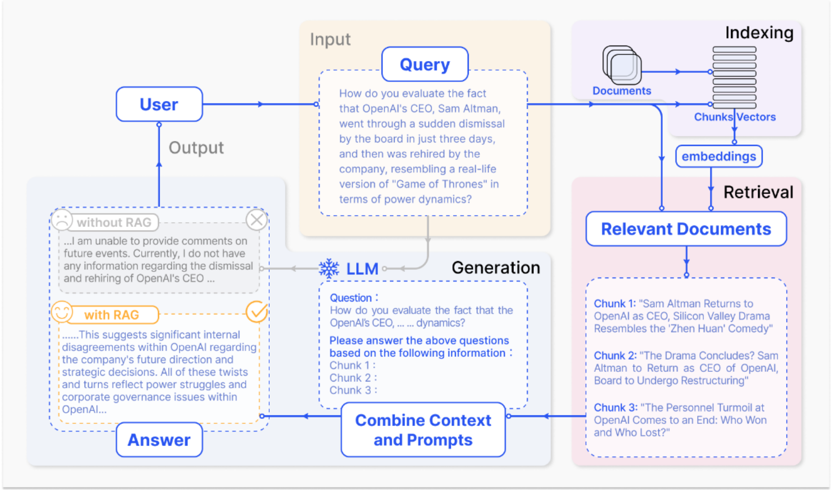


Gao et al. (2023). [Retrieval-Augmented Generation for Large Language Models: A Survey](). arxiv:2312.10997.

How to address factuality issues of LLMs?

Combination with external knowledge

- Detecting and correcting factual mistakes at inference time
- Modularised knowledge-grounded framework
- Retrieval-augmented generation (RAG)



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

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

Overview of Today's Talk

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Determining what parametric knowledge influences a LLM's prediction
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
- **Conclusion**
  - Wrap-up
  - Outlook

Parametric Knowledge and Attribution Methods

- Parametric Knowledge
  - Knowledge acquired during training phase encoded in a LM's weights
  - Our study: change in knowledge acquired during LLM training and task-adaptive training for knowledge-intensive tasks (fact checking, QA, natural language inference)

- Attribution Methods unveil LM's parametric knowledge used to arrive at a prediction
  - Previous methods operate on different levels (instance, neuron)
  - Studied in isolation
  - No consensus as to which methods work best best in which scenarios

We propose a unified evaluation framework that compares two streams of attribution methods, to provide a comprehensive understanding of a LM's inner workings

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.

Parametric Knowledge and Attribution Methods

Instance Attribution (IA) : Find **training instances** that influence the parametric knowledge used by the model

- *Human-interpretable explanation of the model's encoded parametric knowledge*

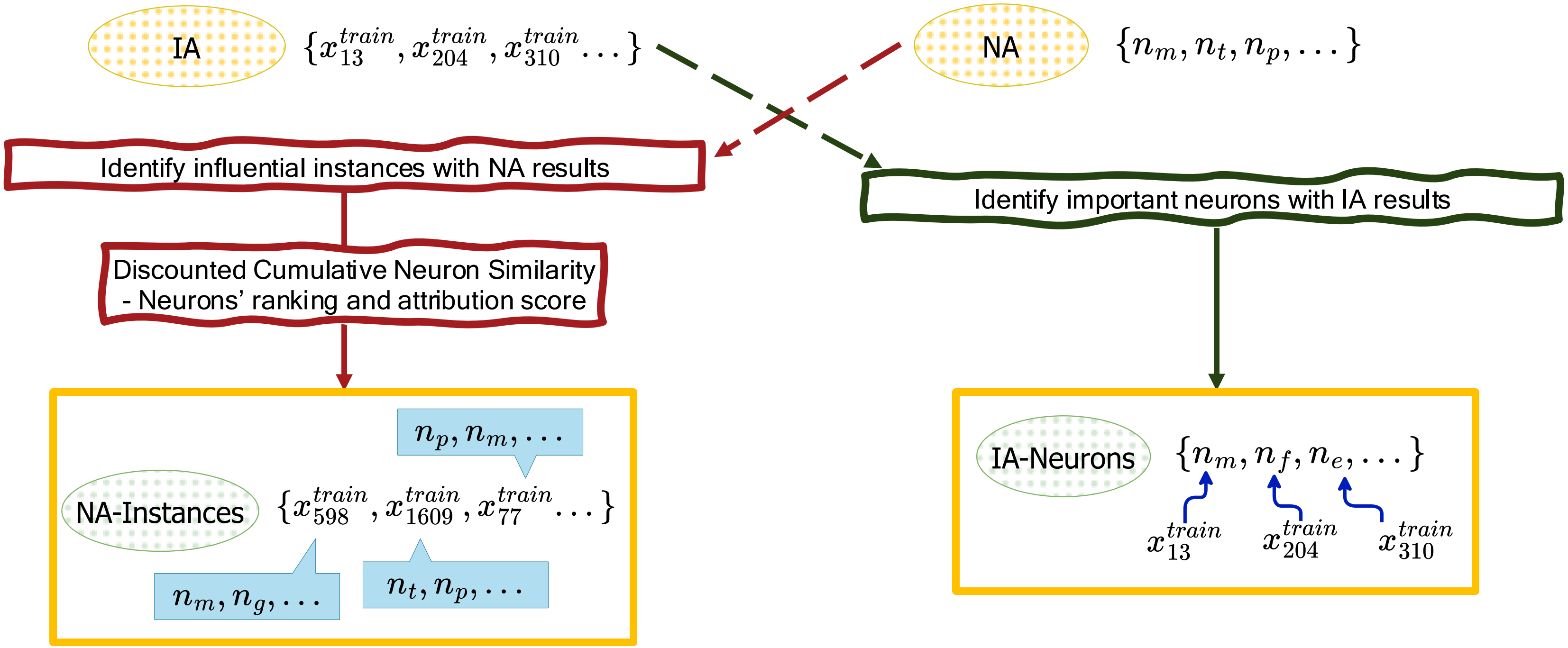
Neuron Attribution (NA) : Locates **specific neurons** that hold the most important parametric knowledge

- *Fine-grained view of which neurons influenced the prediction*

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.

An Evaluation Framework for Attribution Methods

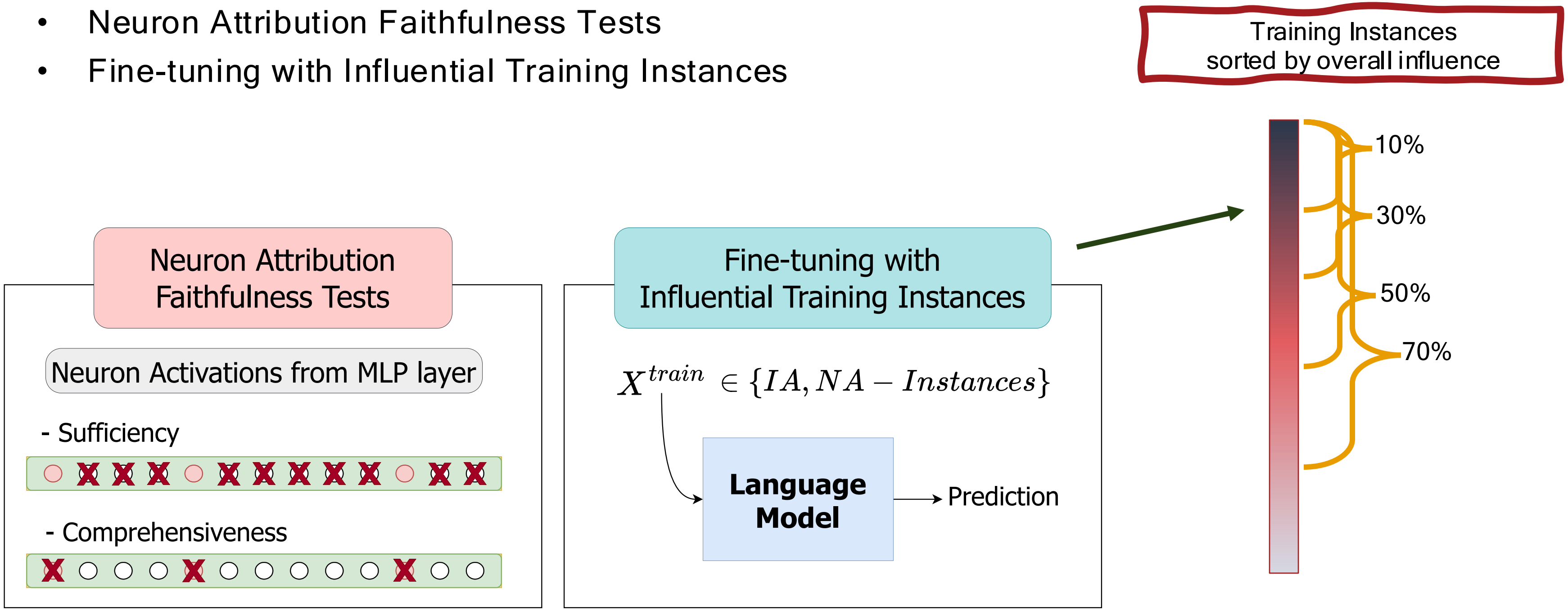
1) Aligning the Results of Attribution Methods



An Evaluation Framework for Attribution Methods

2) Tests

- Neuron Attribution Faithfulness Tests
- Fine-tuning with Influential Training Instances

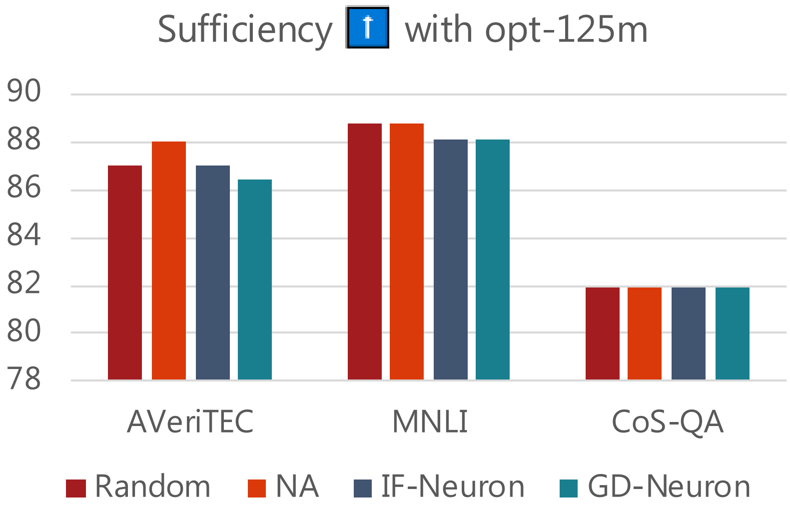


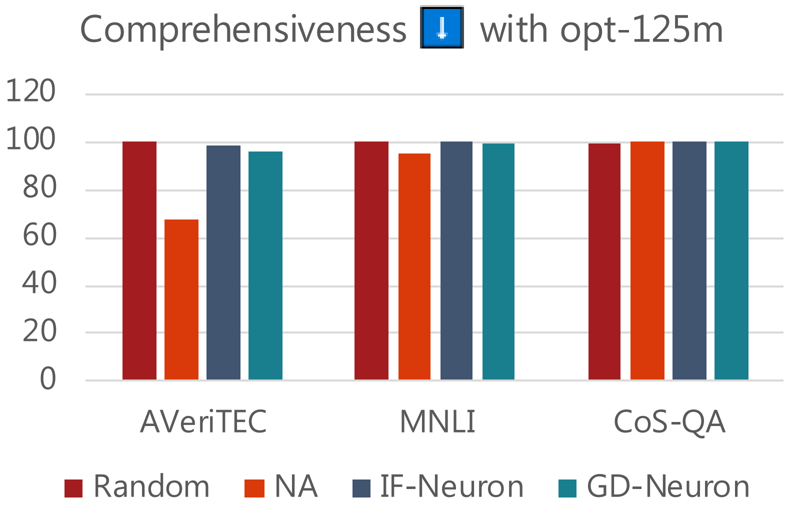
Experimental Set-up

- Instance Attribution
  - Influence Function (IF) (Koh and Liang, 2017), Gradient Similarity (GS) (Charpiat et al., 2019)
- Neuron Attribution
  - The application of Integrated Gradient (Dai et al., 2022)
- Datasets
  - AVeriTeC (Fact-checking) / MNLI (Natural language inference) / Commonsense QA (Question Answering)
- Models
  - opt-125m / Pythia-410m / BLOOM-560m

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.

Neuron Attribution Faithfulness Tests





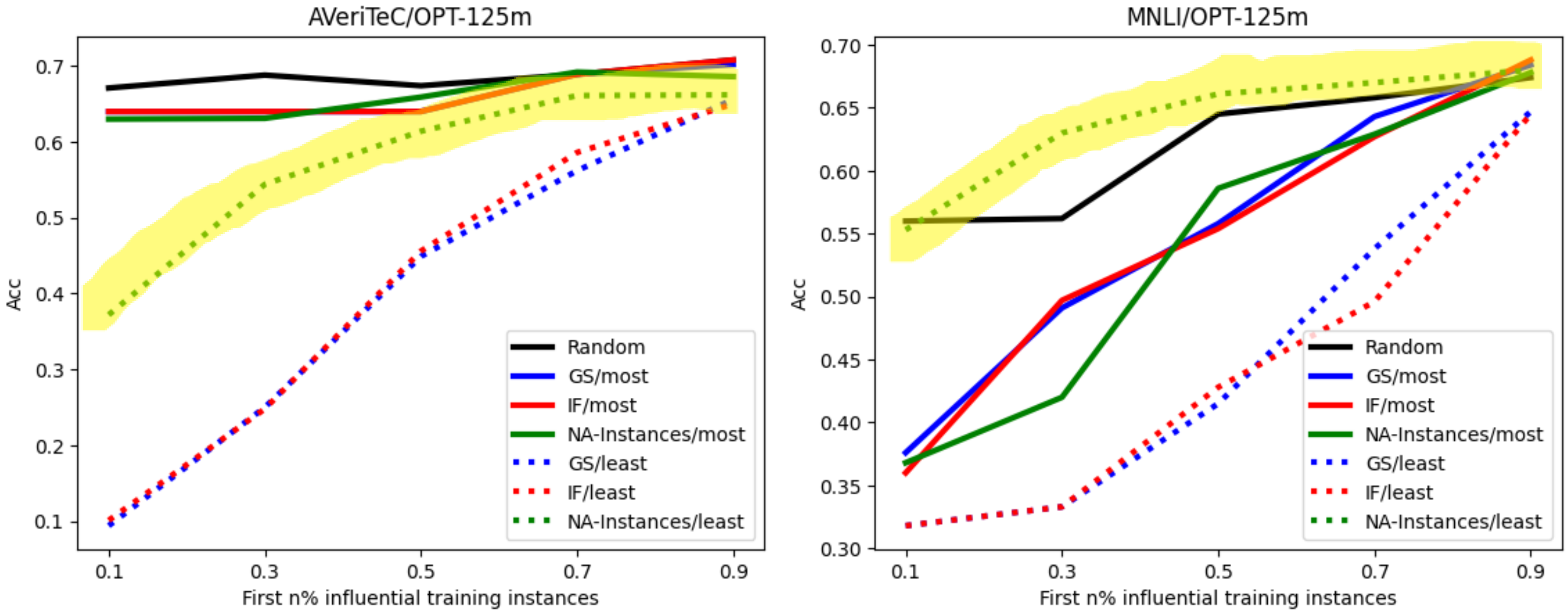
Evaluation metrics

- Random: Randomly select the same number of neurons
- Sufficiency: Only use top-1 important neuron
- Comprehensiveness: Block top-100 neurons

Results

- Marginal differences among methods
- Only 1 neuron can recover prediction with above 70% accuracy
- ➢ Hypothesis: role of attention weights

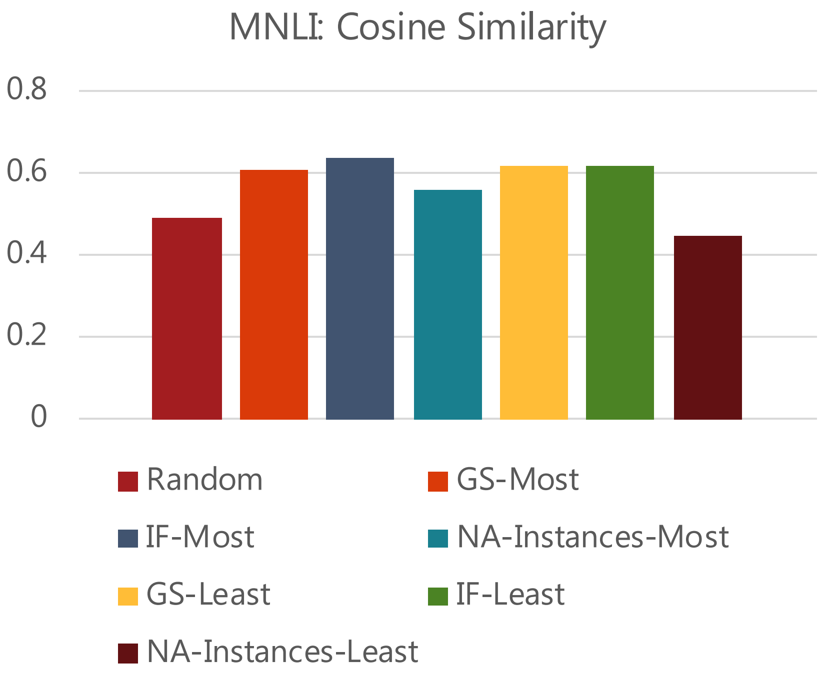
Fine-tuning with Influential Training Instances

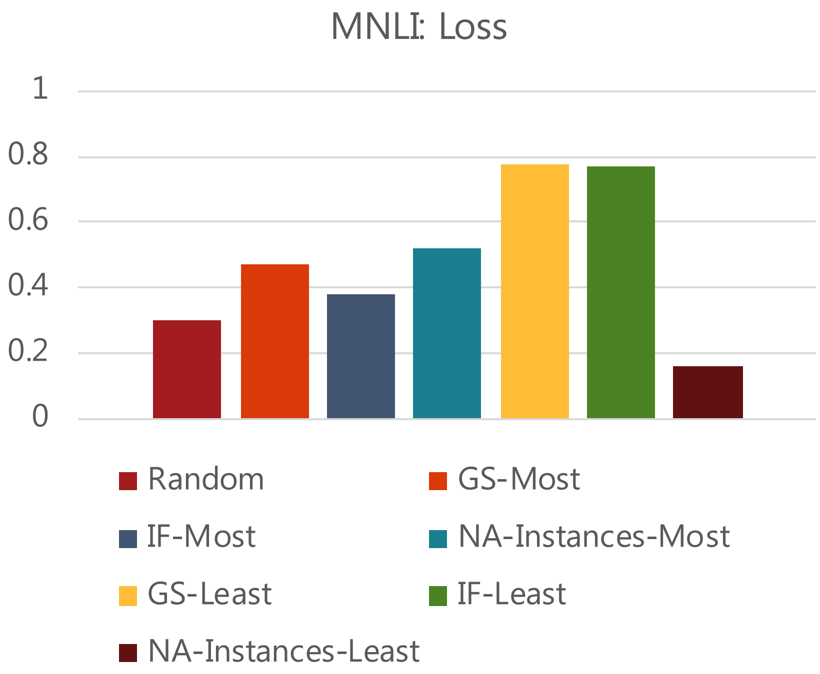


- NA-Instances-Least shows better performance than other least methods
- Counter-intuitive: why would IF-Least perform so well?

➢ Hypothesis: lack of diversity in selected instances

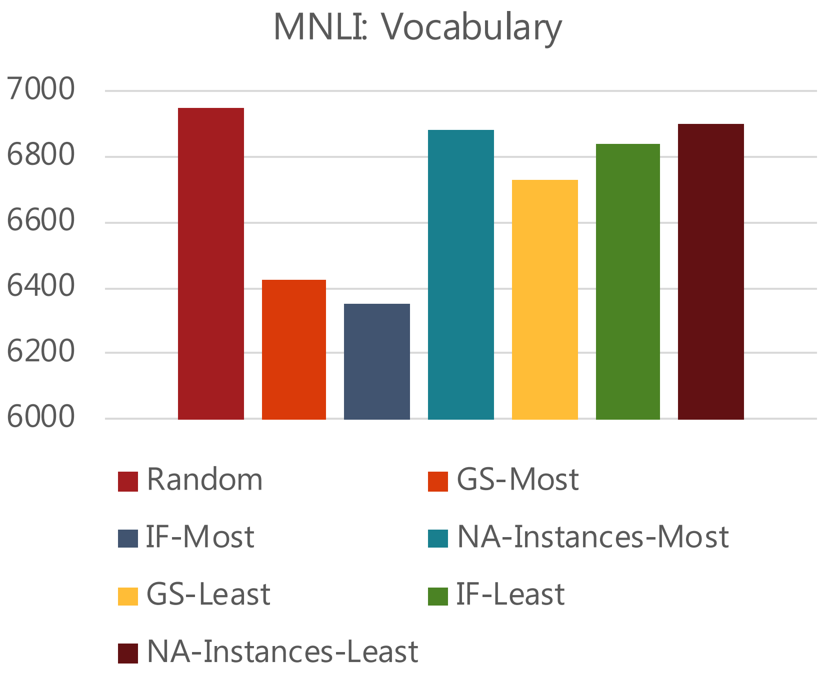
Diversity Analysis on the Group of Influential Training Instances

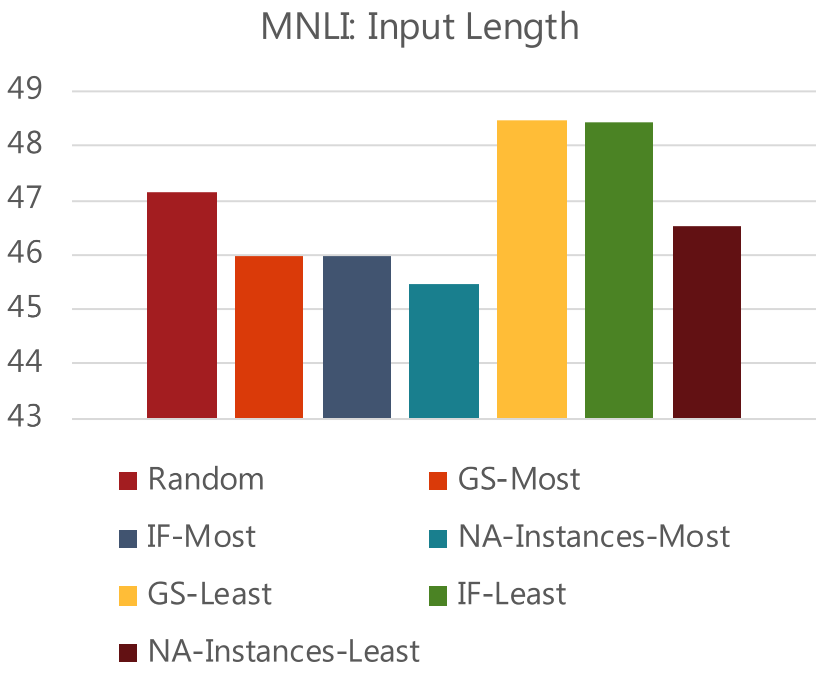




- ➢ NA-Instances-Least results in more diverse instances than Instance Attribution method GS

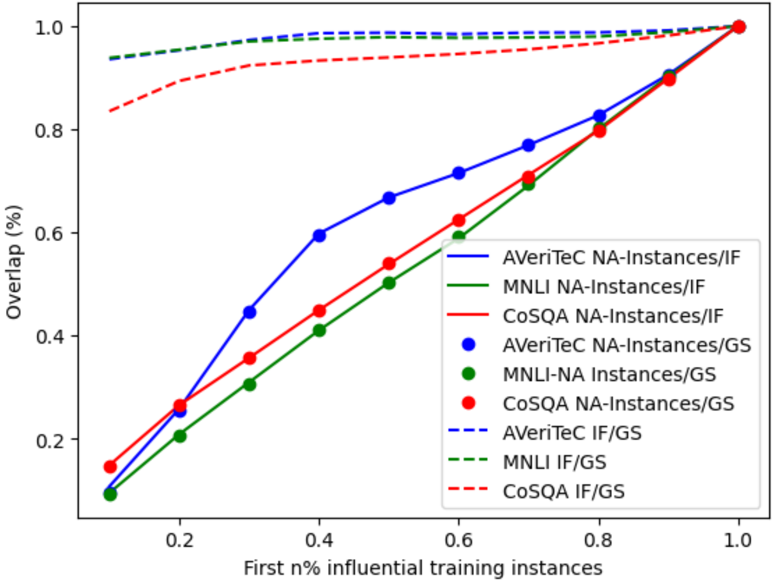
Diversity Analysis on the Group of Influential Training Instances





➢ NA-Instances-Least results in more diverse vocabulary than most other methods

Overlap Analysis of Attribution Methods



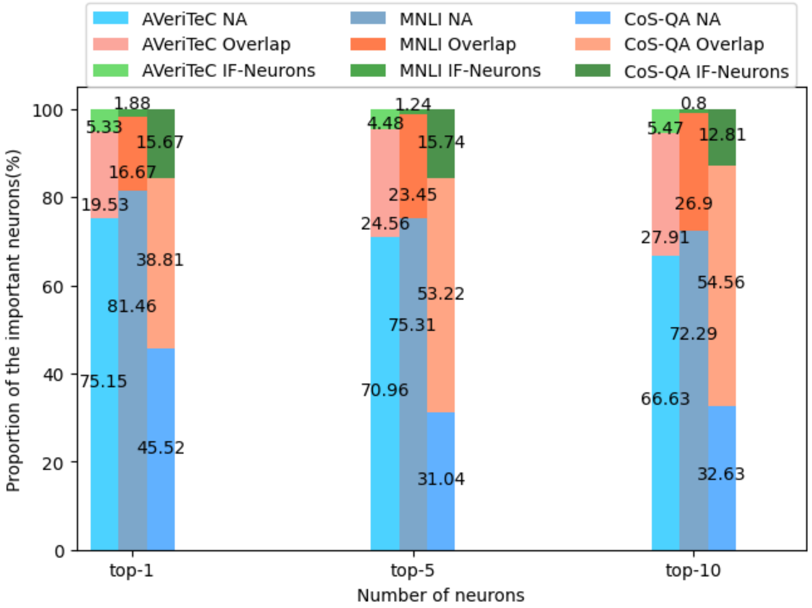
% of training instances at the intersection of the first n% influential instances discovered by a two of the attribution methods ∈ {IF, NA-Instances, and GS}

- High overlap between two instance attribution methods IF and GS

➢ Also explains similar performance on fine-tuning with influential instances

- NA-Instances discovers very different influential instances
- For first 10% of most influential instances discovered by each method, NA-Instances only shares 10% of instances with IA methods IF and GS

Overlap Analysis of Attribution Methods



% of the overlapping top-n important neurons discovered by NA and IF-Neurons

- Proportion of unique important neurons found by NA is higher than those found by IF-Neurons
- ➢ Similar to findings for the diversity of top-n influential training instances

- Most neurons found by IF-Neurons are also discovered by NA
- ➢ NA methods are crucial to reveal the source of the parametric knowledge

Take-Aways: A Unified Framework for Attribution Methods

- We assess the sufficiency and comprehensiveness of the explanations for Instance Attribution and Neuron Attribution with different faithfulness tests

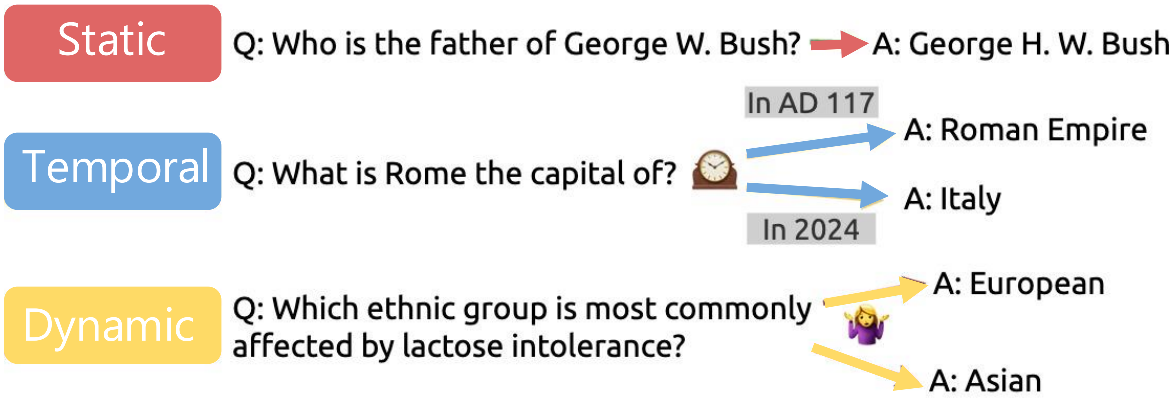
- ➢ Instance Attribution and Neuron Attribution result in **different explanations** about the knowledge responsible for the test prediction
- ➢ Faithfulness tests suggest that **neurons are not sufficient nor comprehensive enough** to fully explain the parametric knowledge used for the test prediction
- ➢ This might be due to the importance of **attention weights** for encoding knowledge

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.

Overview of Today’s Talk

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Determining what parametric knowledge influences a LLM’s prediction
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
- **Conclusion**
  - Wrap-up
  - 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

*

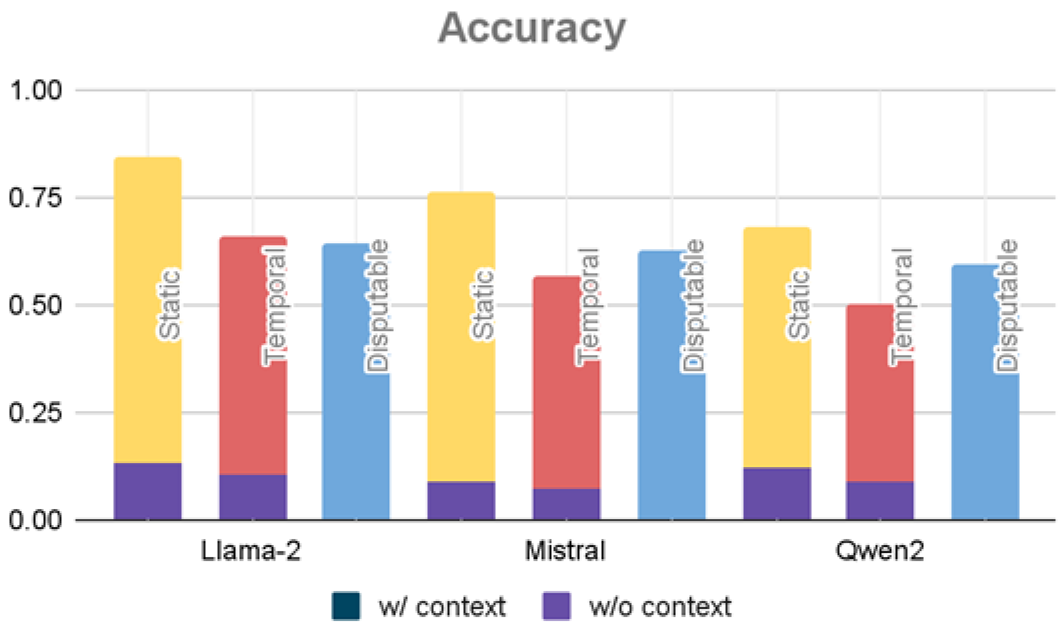
- Wikipedia:Controversial articles

0–9

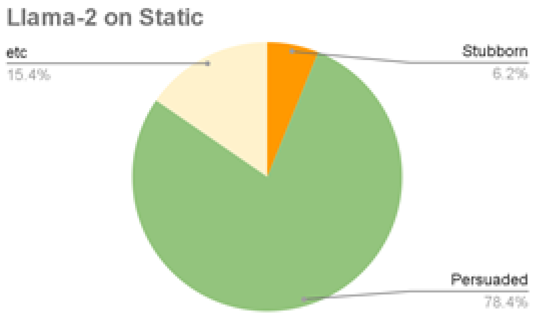
- Talk:2G spectrum case
- Talk:4B movement
- Talk:4chan
- Talk:4chan/Archive 16
- Talk:6ix9ine
- Talk:7 World Trade Center
- Talk:8chan
- Talk:9/11 conspiracy theories
- Talk:9/11 conspiracy theories regarding Jews or Israel
- Talk:10/40 window
- Talk:12 May Karachi riots
- Talk:40 Days for Life
- Talk:44M 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

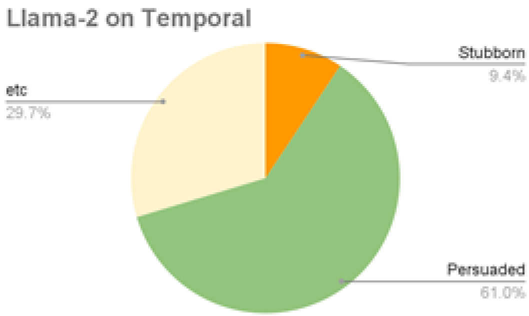
How do LMs perform on the dataset?

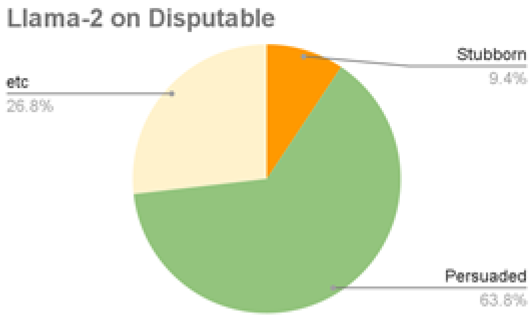
Models perform **best** on static questions, **with and without context.**



How do LMs perform on the dataset?

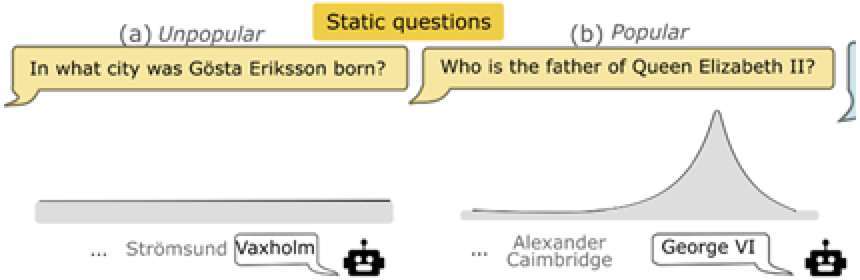




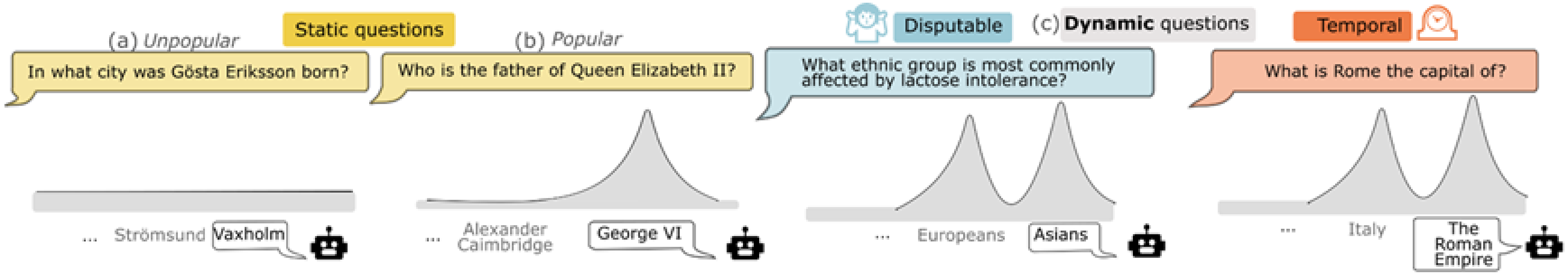


We see more **stubborn instances** in the dynamic partitions
-> Why are **dynamic** facts so **stubborn**?

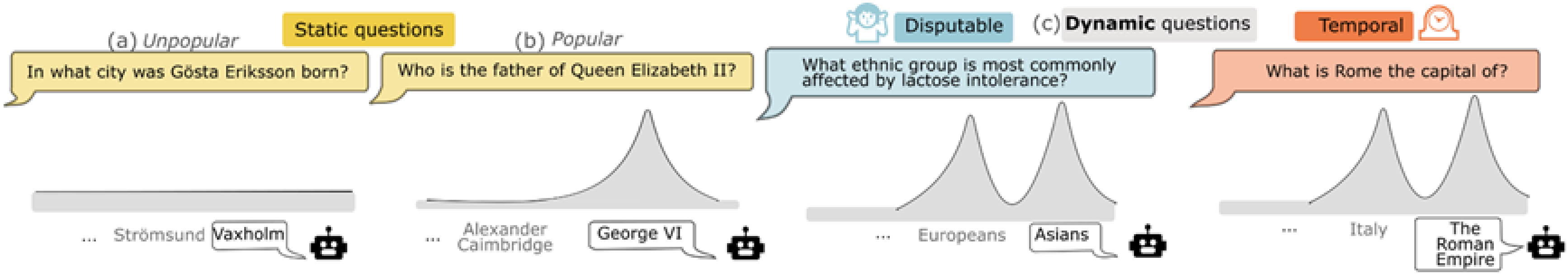
Intra-Memory Conflict in Output Distribution



Intra-Memory Conflict in Output Distribution



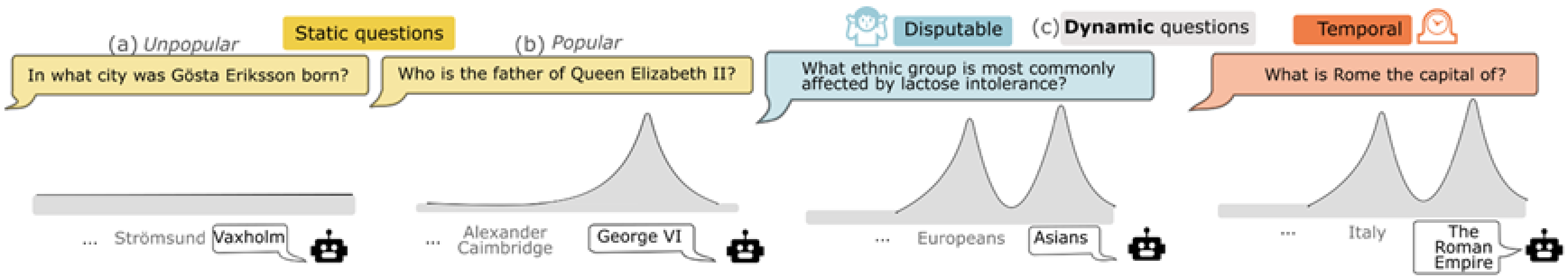
Intra-Memory Conflict in Output Distribution



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

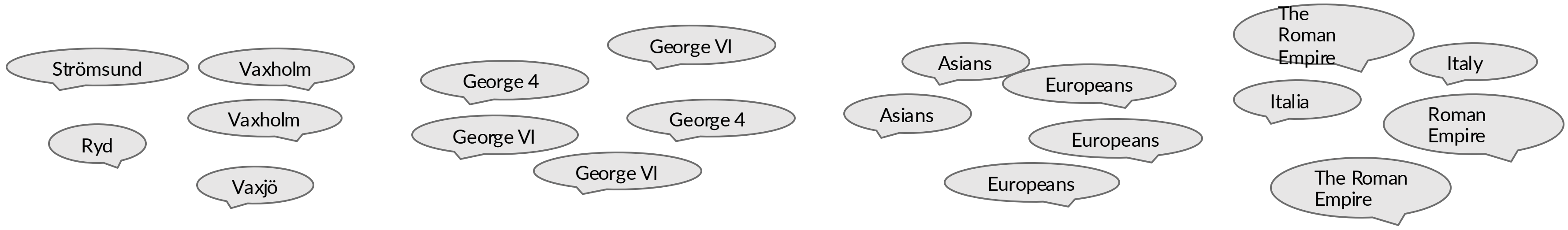
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*

Intra-Memory Conflict in Output Distribution

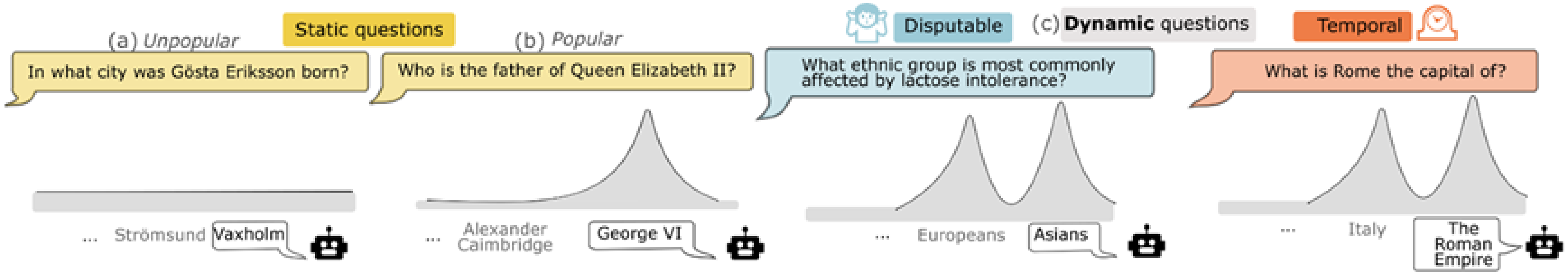


***Dynamic* facts should show greater *entropy* across objects.**

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

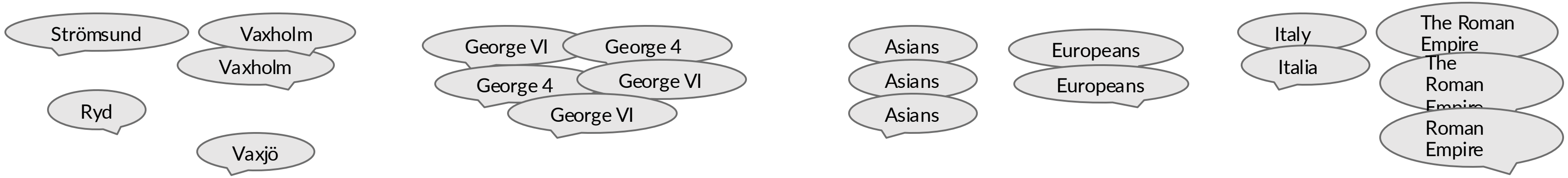


Intra-Memory Conflict in Output Distribution

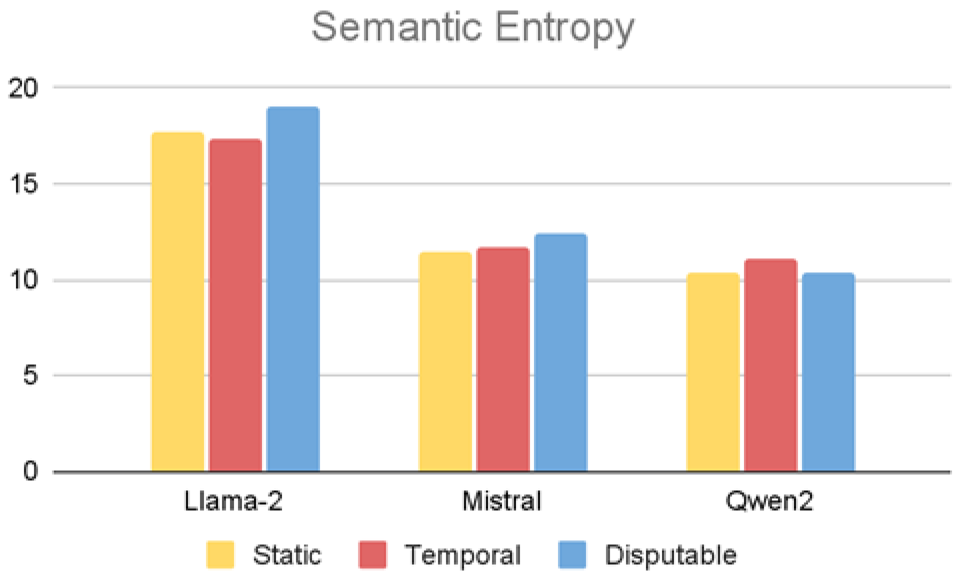


***Dynamic* facts should show greater *entropy* across objects.**

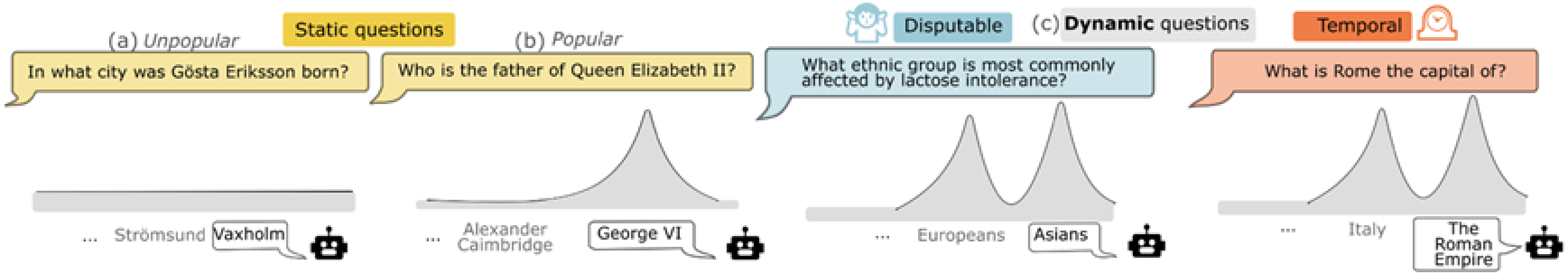
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*



However, this is not always the case

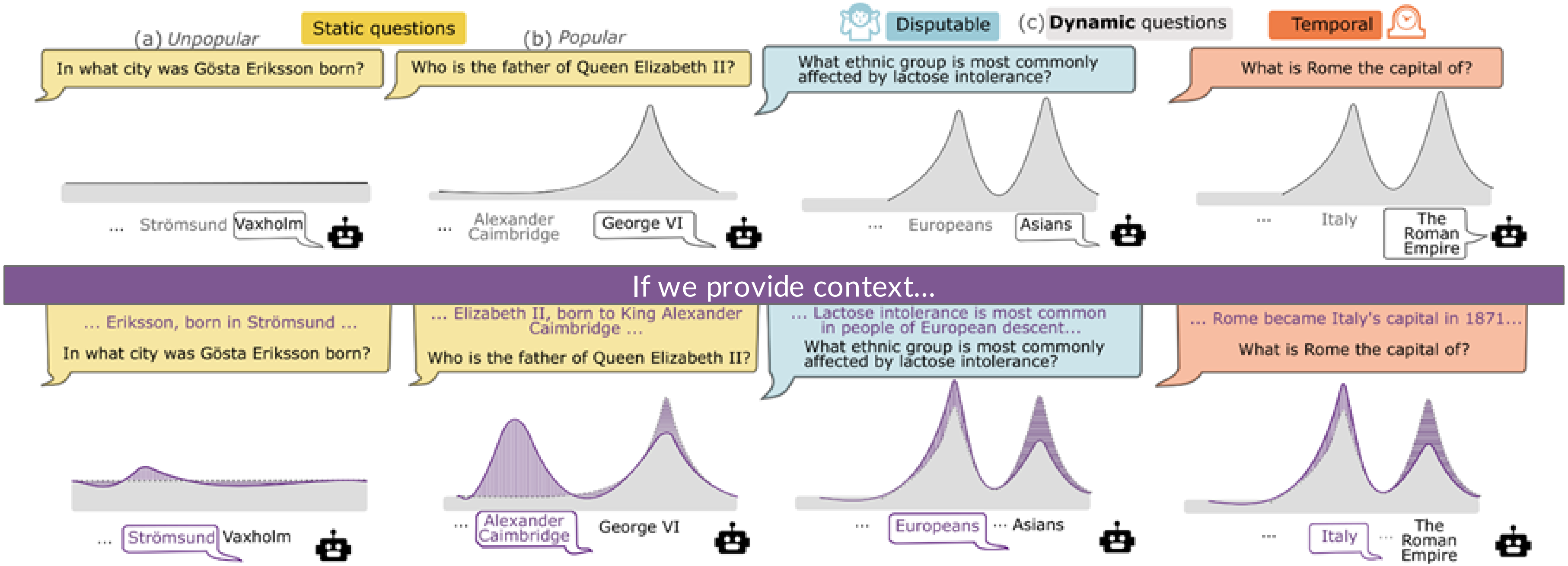


Context-Memory Conflict

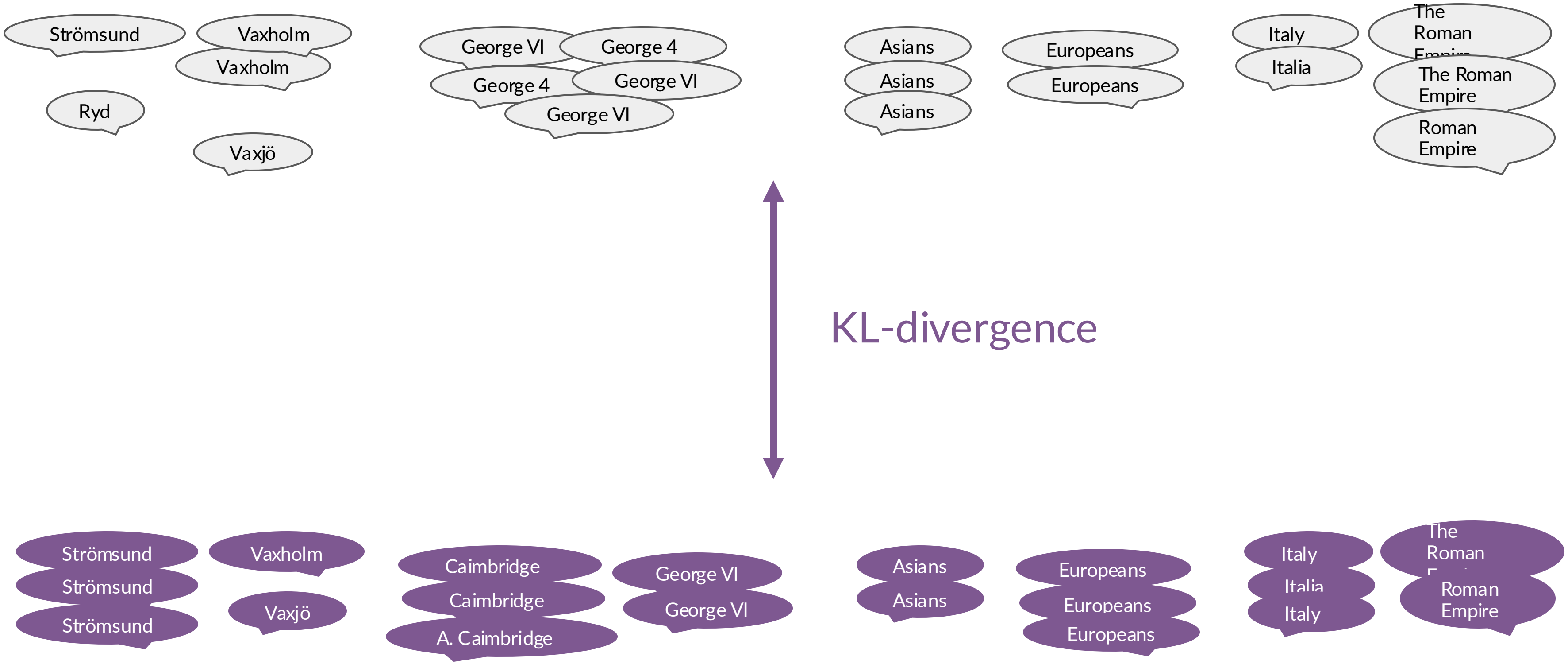


If we provide context…

Context-Memory Conflict

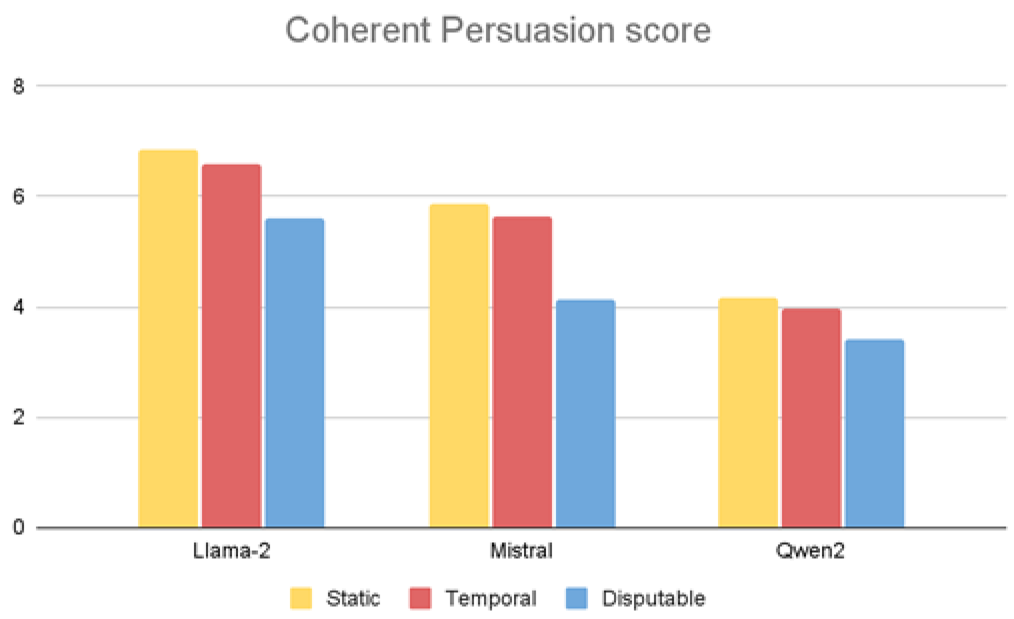


Coherent Persuasion Score



Persuasion Score across Partitions

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



Persuasion Score across Partitions

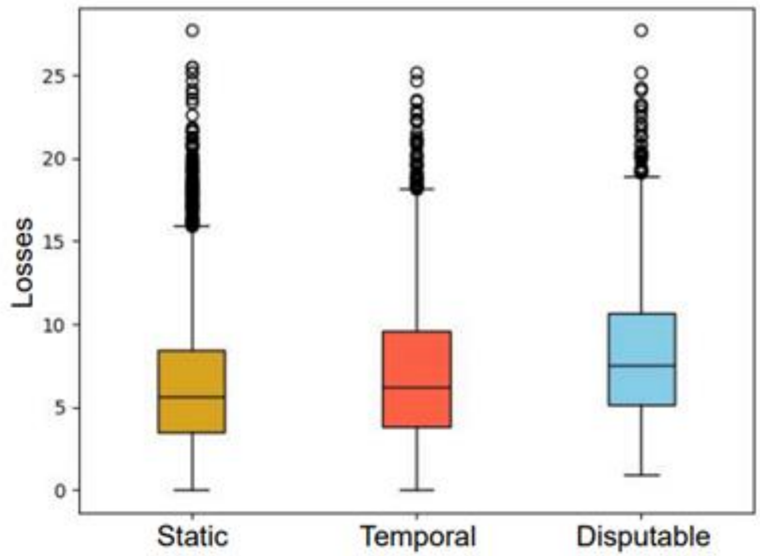
We see the **greatest persuasion score** for the **static dataset.**

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.

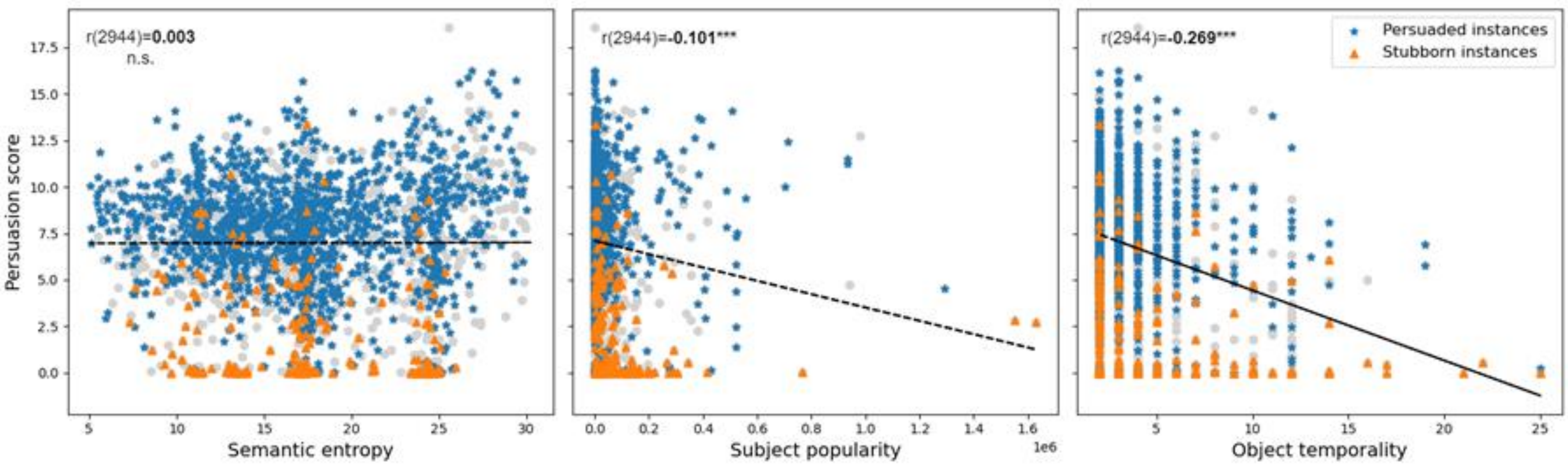
A higher loss indicates greater change required to steer the LM to output the target answer.

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.**

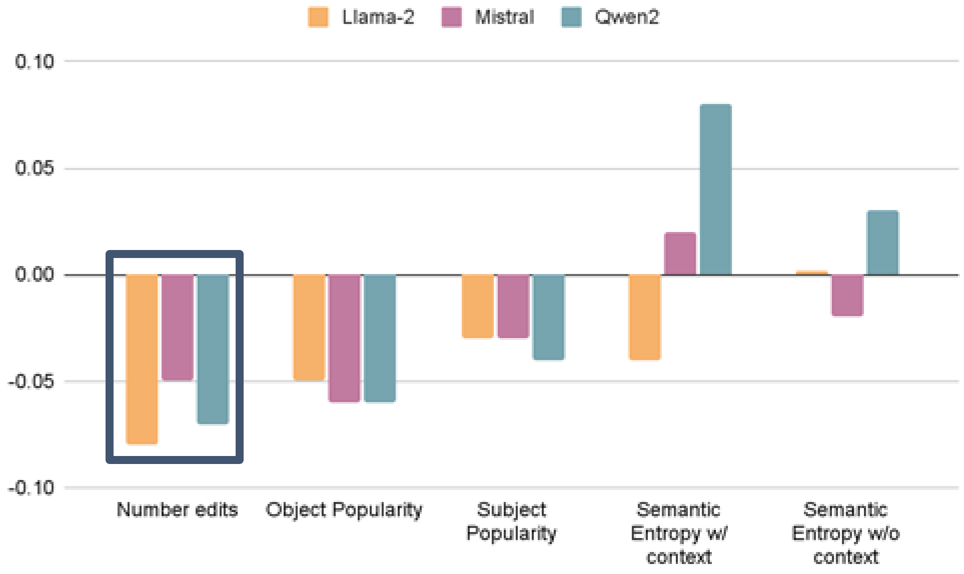
What impacts Persuasion? Correlates with Persuasion

Temporality (number of edits) was the **strongest measured correlate** of model persuasion.



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

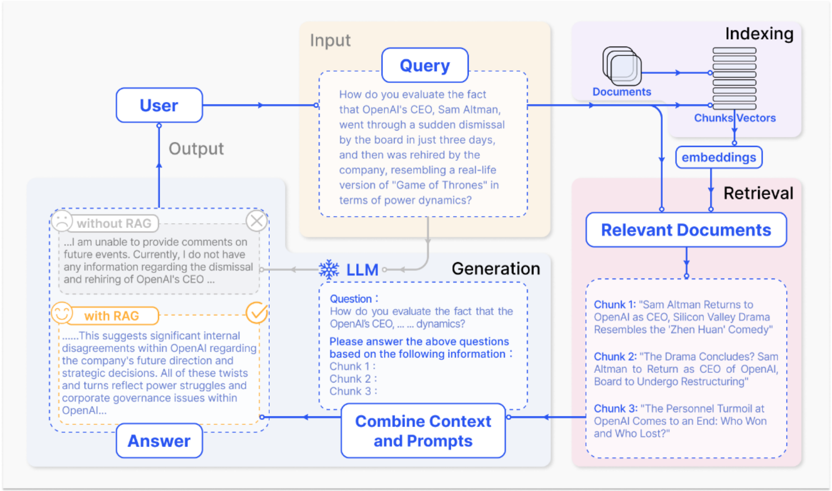
Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Overview of Today’s Talk

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Determining what parametric knowledge influences a LLM’s prediction
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
- **Conclusion**
  - Wrap-up
  - Outlook

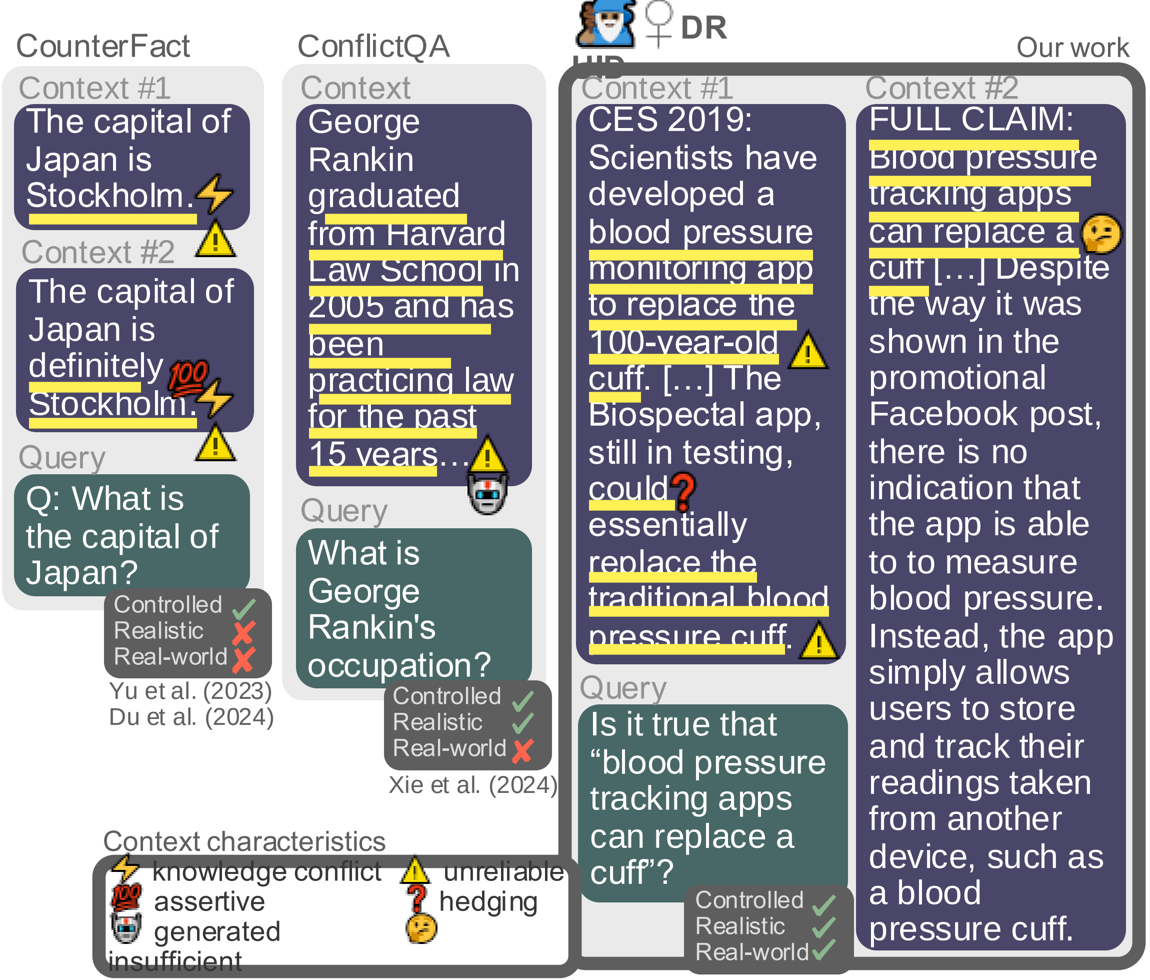
Context Utilisation of Retrieval-Augmented Generation

- 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](). CoRR, abs/2412.17031, December 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](). CoRR, abs/2412.17031, December 2024.

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. CoRR, abs/2412.17031, December 2024.

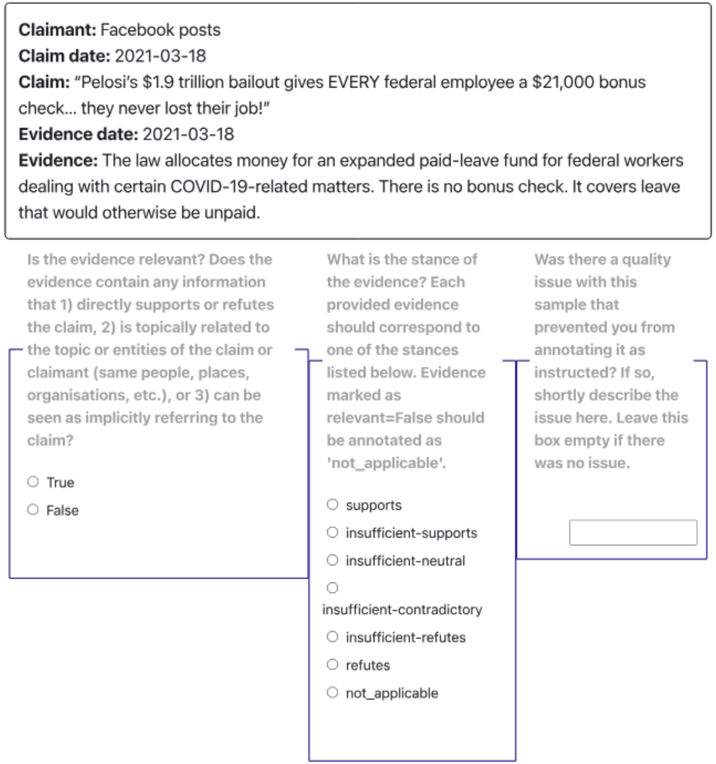
DRUID data selection process

- **Chunk and re-compose**
  - Context compression necessary due to window size constraints
  - Automatically chunk into chunks of 200 words max
  - Get rerank score with Cohere Rerank model
  - Filter out sentences from paragraphs with high overlap, as they only repeat claim
  - Aggregate top 3 chunks

- **Evidence selection**
  - 2 pages published before, 2 after the claim date, gold evidence from fact checking website manually annotated for stance and relevance (DRUID)
  - Rest of evidence pages not annotated, but preserved (DRUID+)

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. CoRR, abs/2412.17031, December 2024.

DRUID data annotation interface


Claimant: Facebook posts
Claim date: 2021-03-18
Claim: "Pelosi's $1.9 trillion bailout gives EVERY federal employee a $21,000 bonus check... they never lost their job!"
Evidence date: 2021-03-18
Evidence: The law allocates money for an expanded paid-leave fund for federal workers dealing with certain COVID-19-related matters. There is no bonus check. It covers leave that would otherwise be unpaid.

Is the evidence relevant? Does the evidence contain any information that 1) directly supports or refutes the claim, 2) is topically related to the topic or entities of the claim or claimant (same people, places, organisations, etc.), or 3) can be seen as implicitly referring to the claim?

- True
- False

What is the stance of the evidence? Each provided evidence should correspond to one of the stances listed below. Evidence marked as relevant=False should be annotated as 'not_applicable'.

- supports
- insufficient-supports
- insufficient-neutral
- insufficient-contradictory
- insufficient-refutes
- refutes
- not_applicable

Was there a quality issue with this sample that prevented you from annotating it as instructed? If so, shortly describe the issue here. Leave this box empty if there was no issue.

Relevant	CounterFact	ConflictQA	DRUID
True	20,000	16,046	5,399
False	0	0	91

Table 8: Evidence relevance for each of the investigated datasets.

Evidence stance	CounterFact	ConflictQA	DRUID
refutes	10,000	8,023	1,760
insufficient	0	0	2,730
-refutes	0	0	557
-contradictory	0	0	410
-neutral	0	0	1,078
-supports	0	0	685
supports	10,000	8,023	909

Table 9: Evidence stance for each of the investigated 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. CoRR, abs/2412.17031, December 2024.

DRUID dataset

Dataset	Claim		Evidence		
	Source	*Type*	*Sufficient*	*Unleaked*	*Retrieved*
FEVER (Thorne et al., 2018)	W	Synthetic	✓	N/A	✓
VitaminC (Schuster et al., 2021)	W	Synthetic	✓	N/A	✓
SciFact (Wadden et al., 2020)	S	Synthetic	✓	N/A	✓
Liar-Plus (Alhindi et al., 2018)	FC	Real	✓	✗	✗
MultiFC (Augenstein et al., 2019)	FC	Real	✗	✗	✓
WatClaimCheck (Khan et al., 2022)	FC	Real	✗	✓	✗
ClaimDecomp (Chen et al., 2022)	FC	Real	✗	✓	✗
Snopes (Hanselowski et al., 2019)	FC	Real	✗	✓	✗
QABrief (Fan et al., 2020)	FC	Real	✗	✓	✗
CHEF (Hu et al., 2022)	FC	Real	✓	✗	✓
AVeriTeC (Schlichtkrull et al., 2024)	FC	Real	✓	✓	✓
Factcheck-Bench (Wang et al., 2024c)	T	Real/Synthetic	✓✗	✓	✓
DRUID	W, FC	Real	✓✗	✓✗	✓

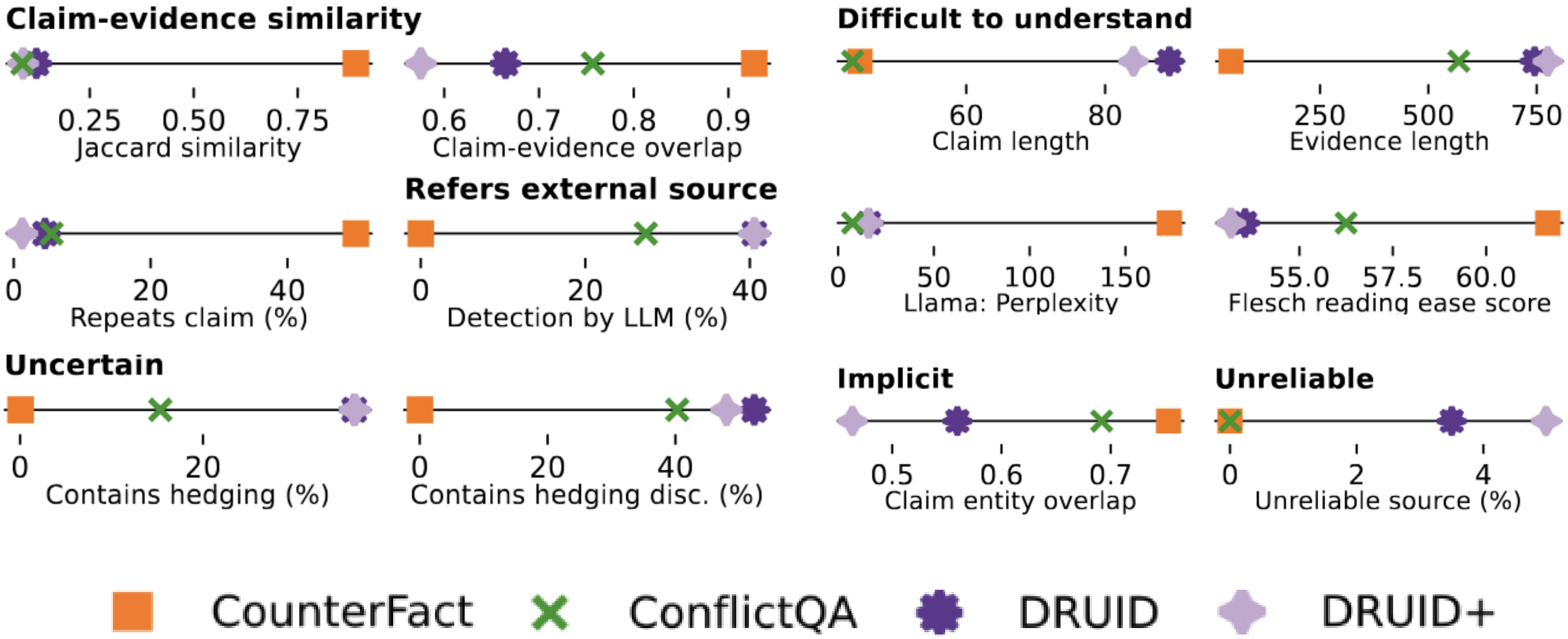
Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. CoRR, abs/2412.17031, December 2024.

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. CoRR, abs/2412.17031, December 2024.

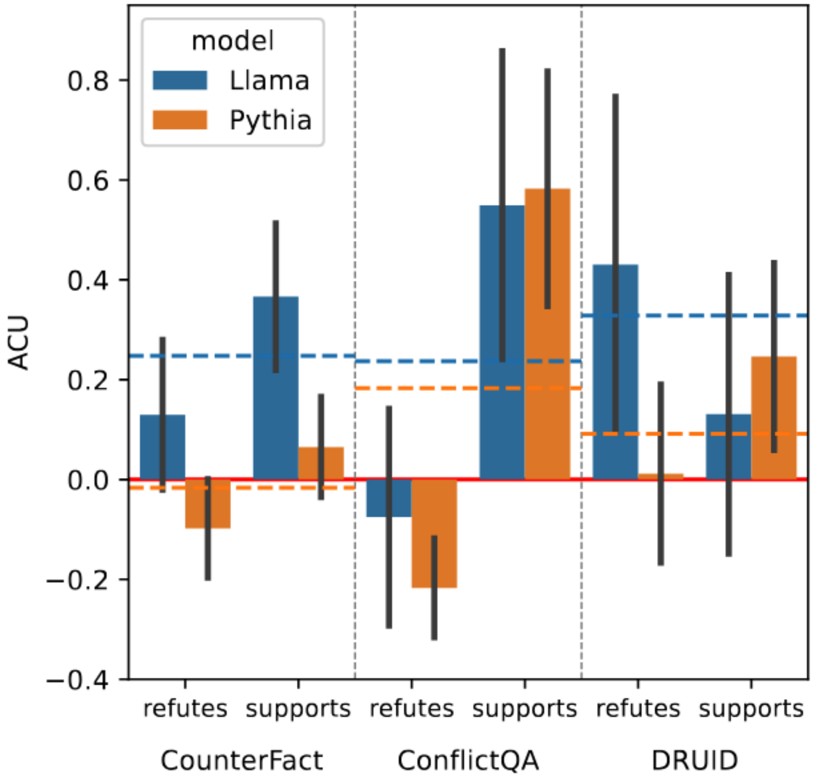
DRUID content characteristics ctd



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](). CoRR, abs/2412.17031, December 2024.

Context utilisation of RAG

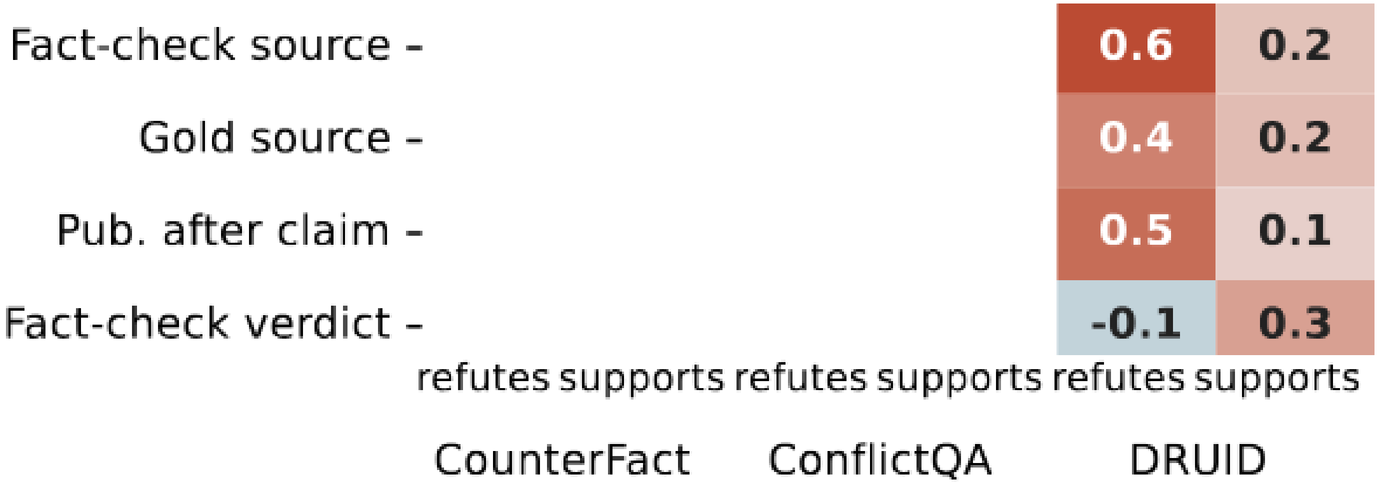
- Context usage (ACU score):
  - Re-scaled difference in salient token probability for difference labels for a claim between settings with vs. without evidence
- **Synthetic** datasets:
  - Over-prefer supporting evidence
  - Context repulsion for refuting evidence
  - Generated automatically -> aligned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



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. CoRR, abs/2412.17031, December 2024.

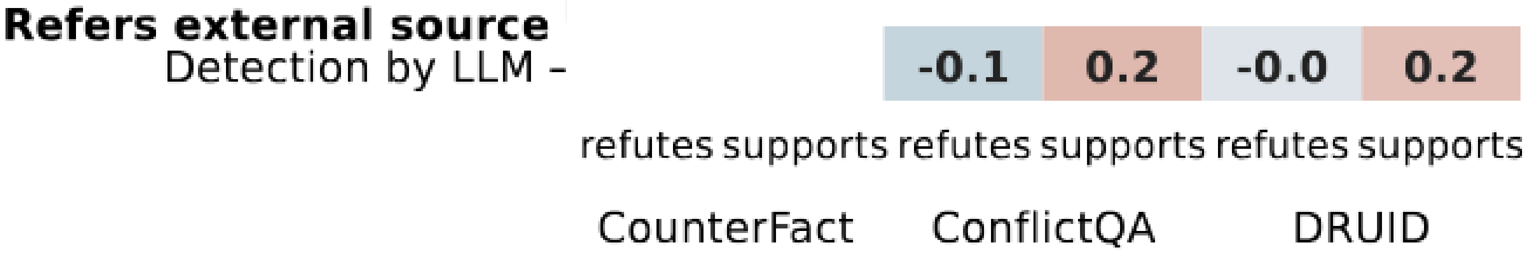
Influence of content characteristics on RAG

- **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’



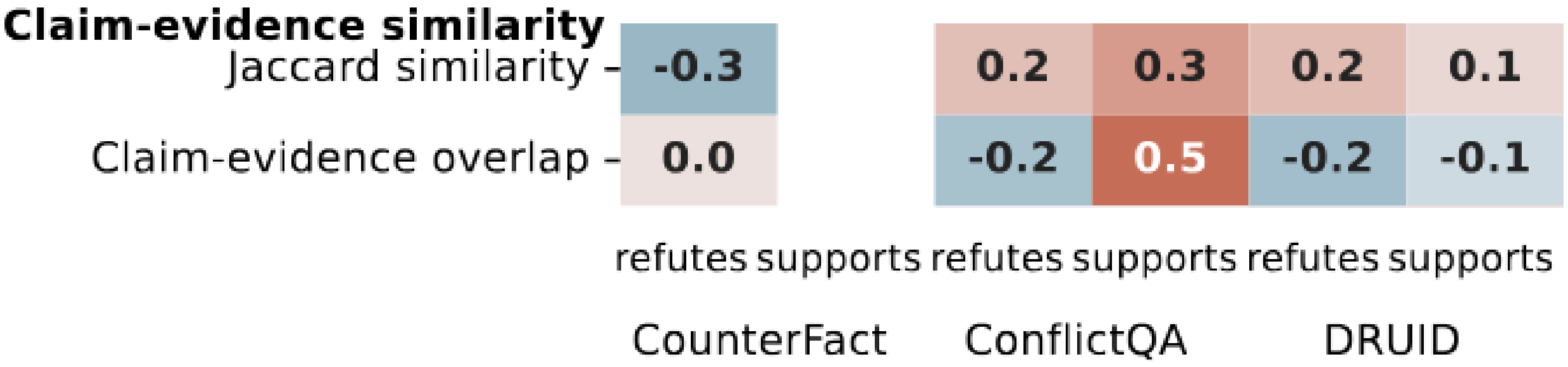
Influence of content characteristics on RAG

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



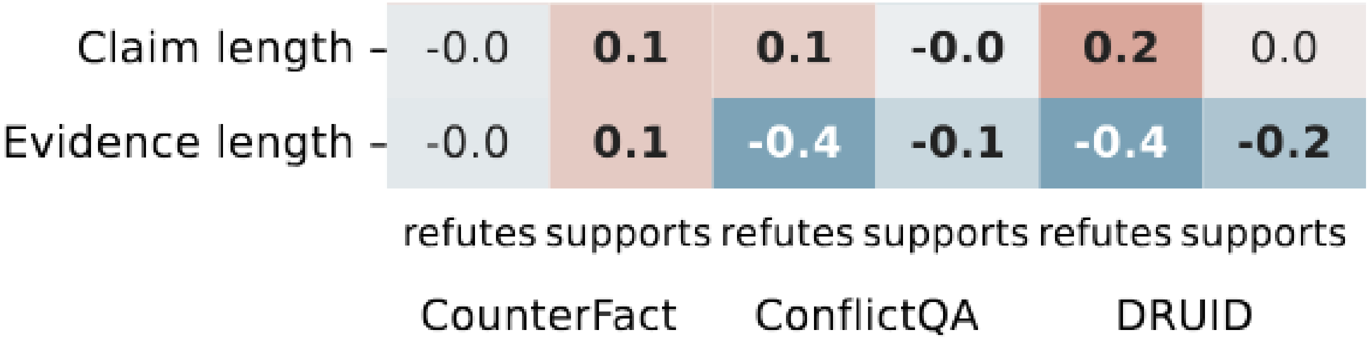
Influence of content characteristics on RAG

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



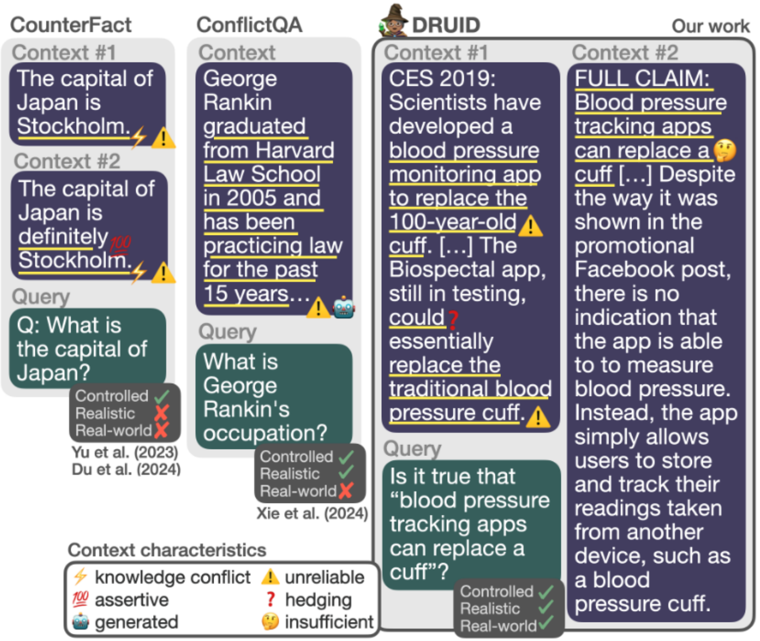
Influence of content characteristics on RAG

- LLMs **less faithful to long contexts**



Take-Aways: Context Utilisation of RAG

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



Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein.** [A Reality Check on Context Utilisation for Retrieval-Augmented Generation](). CoRR, abs/2412.17031, December 2024.

Overview of Today's Talk

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Parametric vs Contextual Knowledge Utilisation of Language Models**
  - Determining what parametric knowledge influences a LLM's prediction
  - Revealing conflicts between parametric and contextual knowledge
  - Determining when or how RAG uses contextual knowledge
- **Conclusion**
  - Wrap-up
  - Outlook

Wrap-Up: Utilisation of Knowledge by LLMs

- How to know **what parametric knowledge influences a LLM's prediction**?
  - Attribution methods can determine knowledge responsible for prediction
  - More work needed to establish their reliability
- How to reveal **conflicts between parametric and contextual knowledge**?
  - Diagnostic test sets with real+counterfactual evidence can reveal how easily a model is persuaded by contextual evidence
  - Models tend to be more stubborn for static than for dynamic facts
- 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

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.

Outlook

- Short and medium-term:
  - Explainability meets RAG
  - Larger-scale comparison of impact of knowledge conflicts
  - Impact of retriever on context use
  - Importance of query context
  - When should context overwrite LLM memory?
- Long-term:
  - LLM scale-up can only achieve so much
  - Revisiting when/how to use LLMs
  - Environmental considerations of LLM usage
  - Next architectural revolution?

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.

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. [Code]

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. [Code], [Data]

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. CoRR, abs/2412.17031, December 2024. [Code], [Data]

CopeNLU Lab

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

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

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

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

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

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

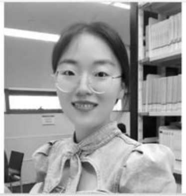
Sarah Masud
Postdoc
Sarah broadly works in the area of computational social systems with a focus on news narrative and hate speech modelling. Her PhD at IIIT-Delhi was supported by fellowships from Google and PMRF.

Arnav Arora
PhD Student
Arnav's research interests include equitable ML, mitigating online harms, and the intersection of NLP and Computational Social Science.

Erik Arakelyan
PhD Student
Erik's main research interests are question answering and explainability.

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 incluc explainability, fact-checking, and question answering.

Siddhesh Pawar
PhD Student
Siddhesh Pawar's research intere include multilingual models, fairn and accountability in NLP system

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.

Na Min An
PhD Intern
Na Min An's research interests are explainability, multimodal systems, and human-centered AI.

Thank you for your attention! Questions?