Detecting Factual Errors of Large Language Models

Isabelle Augenstein*

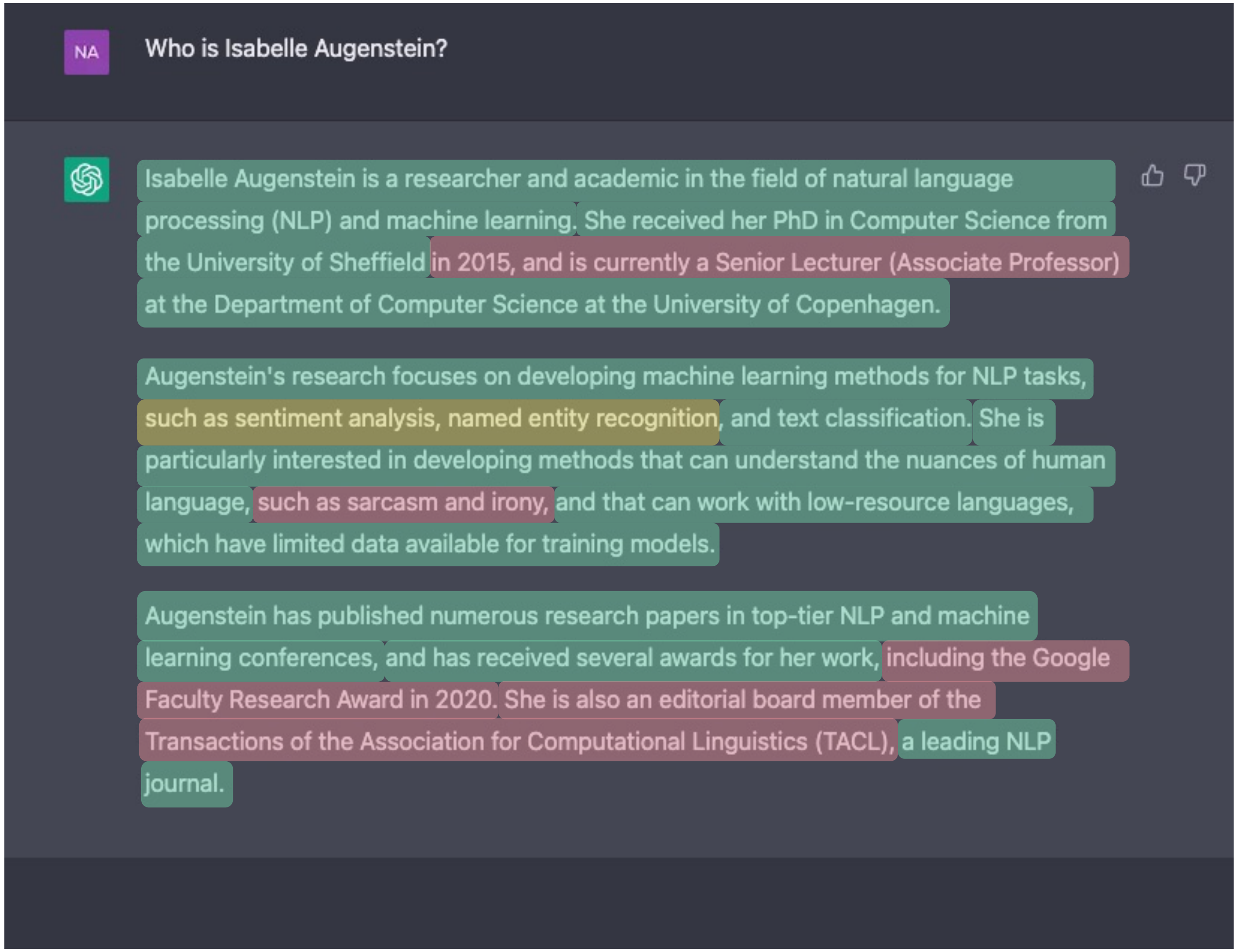
Sprogteknologisk Konference 2024
28 November 2024

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

*Partial credit for slides: Haeun Yu, Sara Vera Marjanovic

UNIVERSITY OF COPENHAGEN

Factuality Challenges of Large Language Models



Factuality Challenges of Large Language Models

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

Factuality Challenges of Large Language Models

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

Overview of Today's Talk

- **Introduction**
  - Factuality Challenges of Large Language Models
- **Post-Hoc Detection and Correction of Factual Errors**
  - Fact Checking and Correction of Machine-Generated Content
- **Probing the Parametric Knowledge of Language Models**
  - A Unified Framework for Input Feature Attribution Methods
  - Detecting Knowledge Conflicts of Language Models
- **Conclusion**
  - Wrap-up
  - Outlook

The Conventional Fact Checking Pipeline

Claim Check-Worthiness Detection

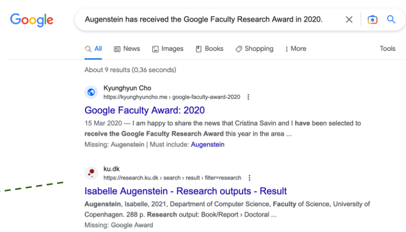
“Augenstein has published … and has received several awards, including the Google Faculty Research Award in 2020.”

not check-worthy

check-worthy

Evidence Document Retrieval

”Augenstein has received the Google Faculty Research Award in 2020.”



Stance Detection / Textual Entailment

”Augenstein has received the Google Faculty Research Award in 2020.; ”Past programs: Faculty research awards program (2005-2019), Focused research awards (2009-2020), …”

positive

neutral

negative

Veracity Prediction

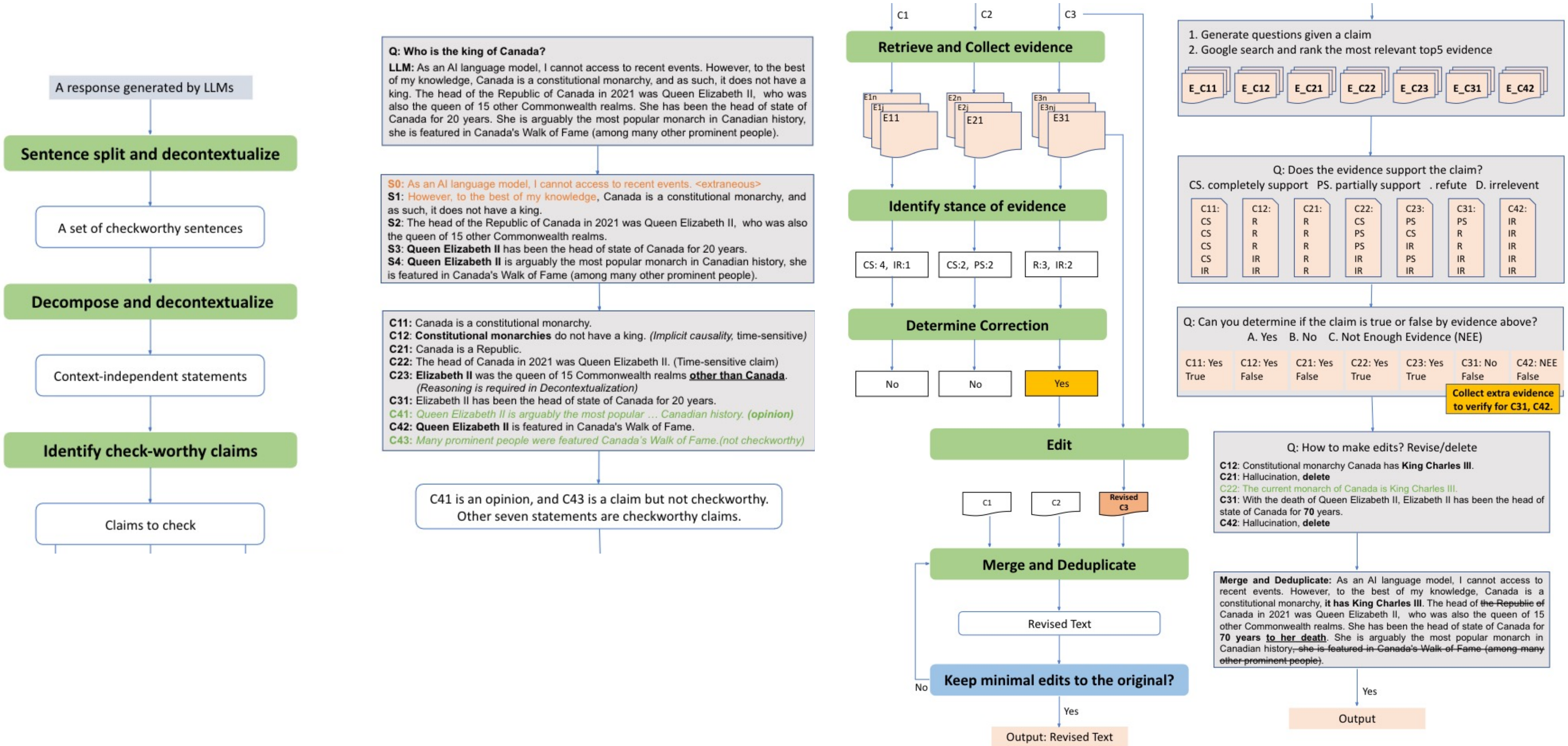
“Augenstein has published … and has received several awards, including the Google Faculty Research Award in 2020.”

true

not enough info

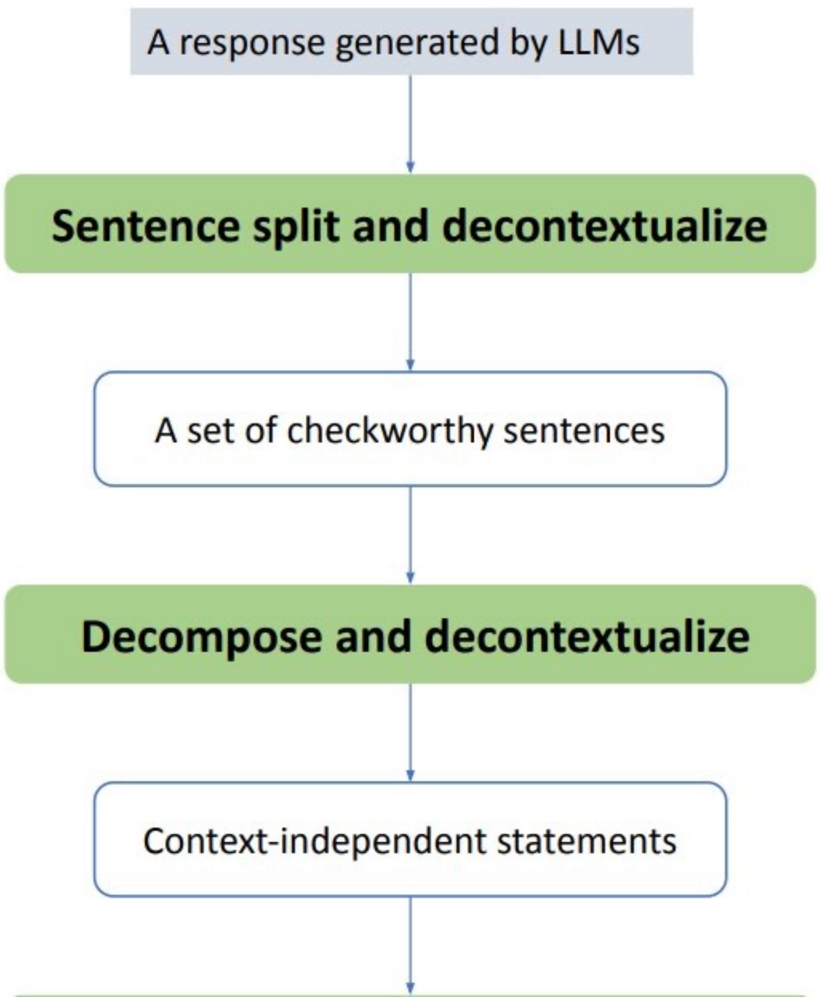
false

Fact Checking and Correction of Machine-Generated Misinformation



Yuxia Wang et al. [Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output]. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing ([EMNLP 2024]), November 2024.

Fact Checking and Correction of Machine-Generated Misinformation



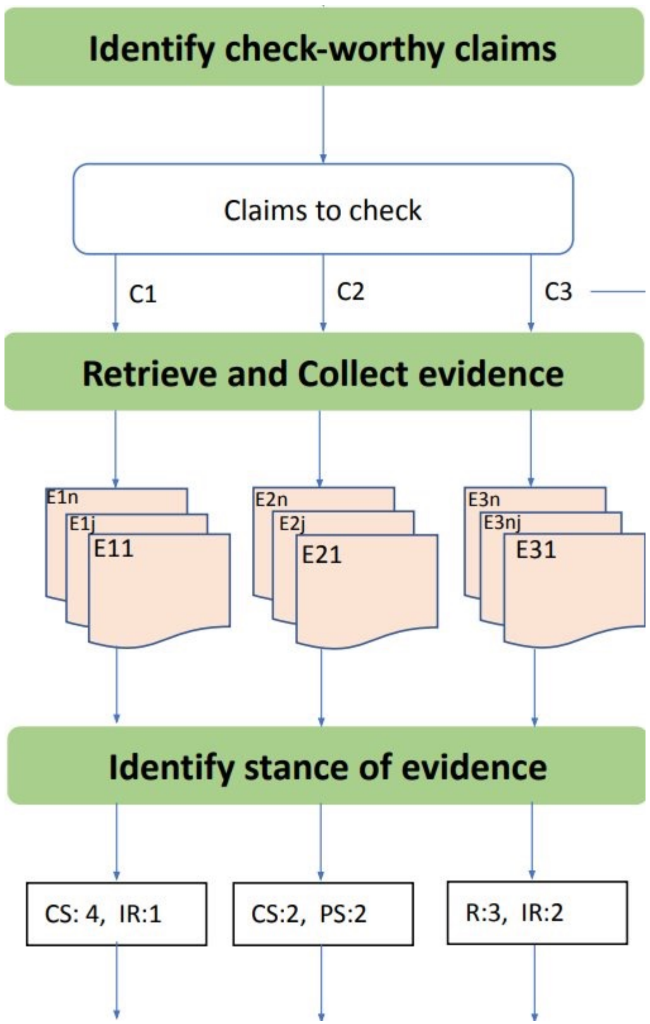
Q: Who is the king of Canada?

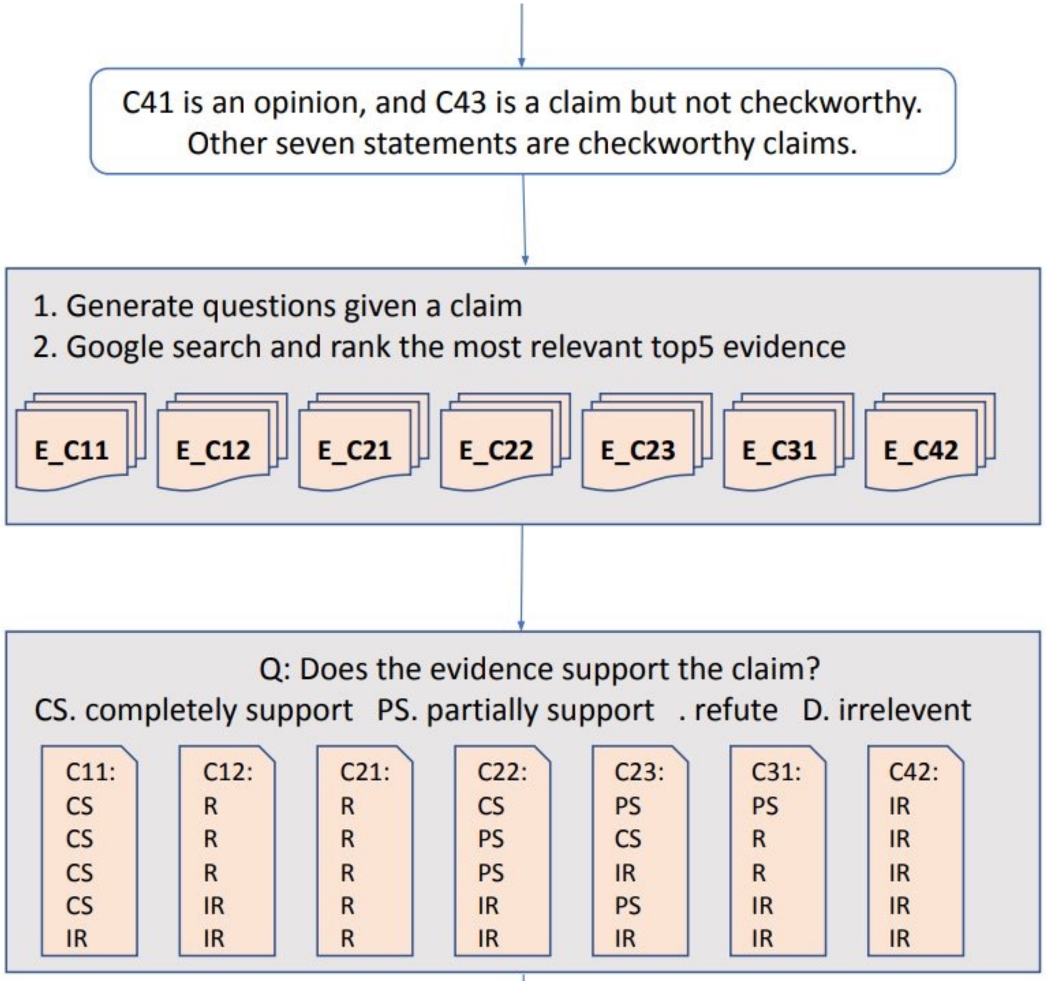
LLM: As an AI language model, I cannot access to recent events. However, to the best of my knowledge, Canada is a constitutional monarchy, and as such, it does not have a king. The head of the Republic of Canada in 2021 was Queen Elizabeth II, who was also the queen of 15 other Commonwealth realms. She has been the head of state of Canada for 20 years. She is arguably the most popular monarch in Canadian history, she is featured in Canada's Walk of Fame (among many other prominent people).

S0: As an AI language model, I cannot access to recent events. <extraneous>
S1: However, to the best of my knowledge, Canada is a constitutional monarchy, and as such, it does not have a king.
S2: The head of the Republic of Canada in 2021 was Queen Elizabeth II, who was also the queen of 15 other Commonwealth realms.
S3: **Queen Elizabeth II** has been the head of state of Canada for 20 years.
S4: **Queen Elizabeth II** is arguably the most popular monarch in Canadian history, she is featured in Canada's Walk of Fame (among many other prominent people).

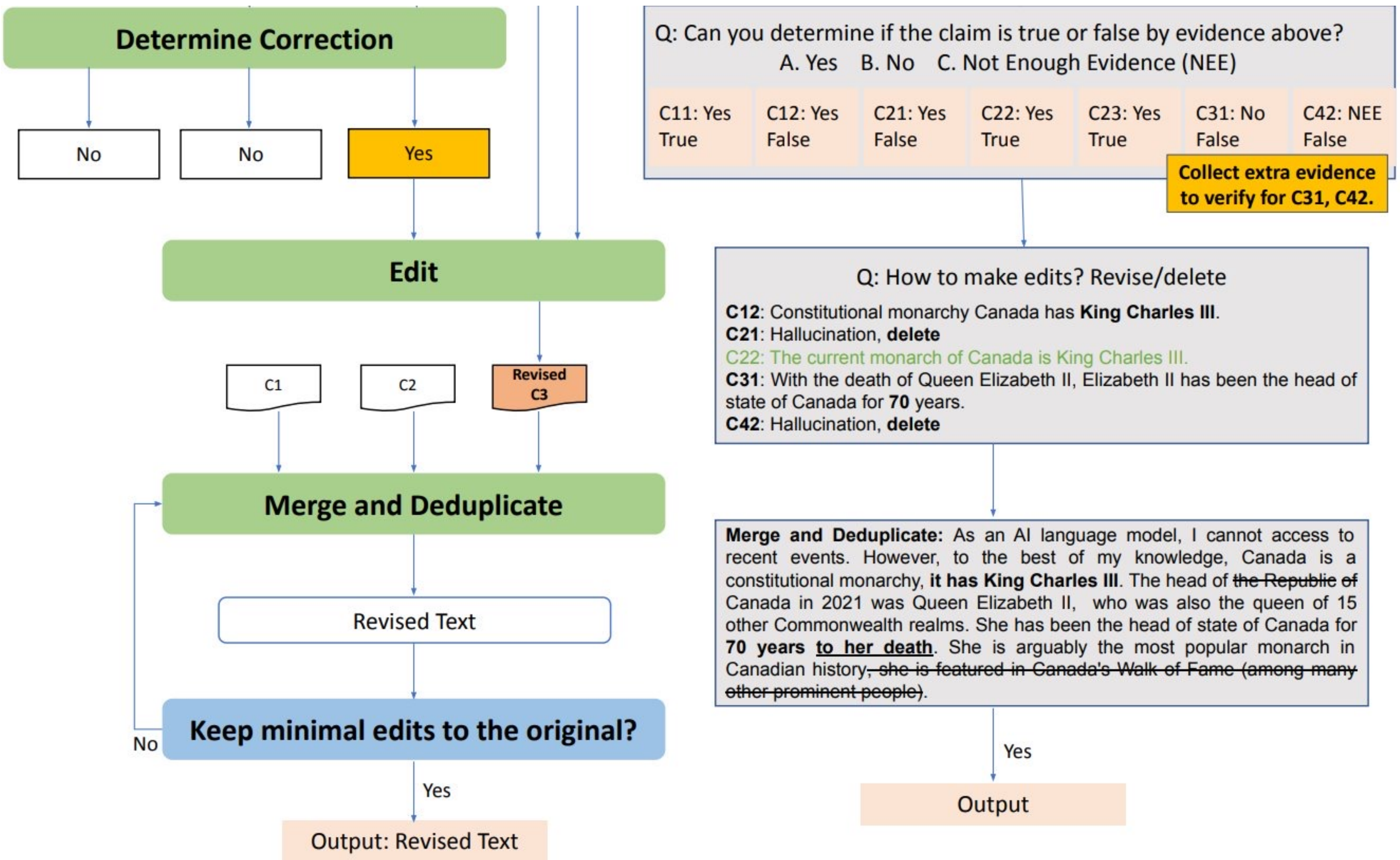
C11: Canada is a constitutional monarchy.
C12: **Constitutional monarchies** do not have a king. *(Implicit causality,* time-sensitive*)*
C21: Canada is a Republic.
C22: The head of Canada in 2021 was Queen Elizabeth II. (Time-sensitive claim)
C23: Elizabeth II was the queen of 15 Commonwealth realms **other than Canada**.
(Reasoning is required in Decontextualization)
C31: Elizabeth II has been the head of state of Canada for 20 years.
C41: *Queen Elizabeth II is arguably the most popular … Canadian history.* ***(opinion)***
C42: Queen Elizabeth II is featured in Canada's Walk of Fame.
C43: *Many prominent people were featured Canada's Walk of Fame.(not checkworthy)*

Fact Checking and Correction of Machine-Generated Misinformation

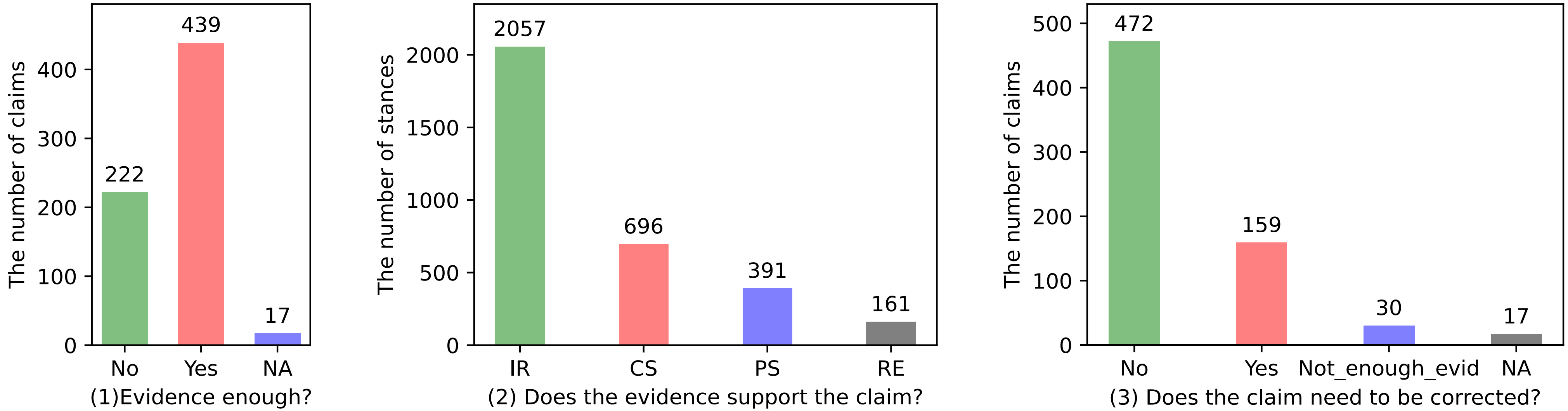




Fact Checking and Correction of Machine-Generated Misinformation



Data Collection of Machine-Generated Misinformation



Claim analysis: (1) whether raters can determine the factuality of a claim depending on the automatically-collected evidence (*Yes/No*); (2) does the evidence support the claim (*CP*: completely support, *PS*: partially support, *RE*: refute, *IR*: irrelevant); (3) does the claim need to be corrected. NA (17) refers to 16 opinion-claims + 1 *not-a-claim*.

Yuxia Wang et al. Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Evaluation of Automatic Factcheck-GPT Pipeline

Task	Method	Acc	Prec	Recall	F1-macro
1	Always-checkworthy	0.891	0.445	0.500	0.471
1	ChatGPT	0.814	0.637	0.740	0.660
2	Always-checkworthy	0.975	0.325	0.333	0.329
2	ChatGPT	0.932	0.314	0.534	0.319

Table 3: **Checkworthiness** detection by majority guess: Always-checkworthy vs. ChatGPT zero-shot prompt. *average*="macro" is used in precision (Pred), recall and F1 calculation.

Method	Acc	Prec	Recall	F1-macro
Four-label space				
Random guess	0.255	0.258	0.264	0.215
LLaMA2-zeroshot	0.202	0.324	0.280	0.155
ChatGPT-zeroshot	0.365	0.402	0.439	0.332
Three-label space				
ChatGPT-zeroshot	0.567	0.506	0.588	0.483
LLaMA2-zeroshot	0.401	0.407	0.384	0.299
RoBERTa-large-mnli	**0.607**	**0.536**	**0.609**	**0.512**

Table 4: **Stance** detection by ChatGPT and LLaMA2 zero-shot prompt. Three-label space merges complete and partial support into one.

Yuxia Wang et al. Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Evaluation of Automatic Factcheck-GPT Pipeline

Verifier	Source	Label = True			Label = False		
		Prec	Recall	F1	Prec	Recall	F1
Random	NA	0.79	0.43	0.56	0.18	0.52	0.27
Always True	NA	0.81	1.00	0.88	0.00	0.00	0.00
Always False	NA	0.00	0.00	0.00	0.19	1.00	0.33
Inst-LLAMA	Wiki	0.87	0.74	0.80	0.34	0.56	0.42
Inst-LLAMA	Web	0.88	0.80	0.84	0.40	0.56	0.47
GPT-3.5-Turbo	Wiki	0.87	0.67	0.76	0.31	0.60	0.41
GPT-3.5-Turbo	Web	0.89	0.74	0.81	0.37	0.62	0.46
Perplexity.ai	Web	**0.93**	**0.73**	**0.83**	0.40	0.76	0.53
Factcheck-GPT	Web	0.90	0.71	0.79	**0.52**	**0.80**	**0.63**

Table 5: **Verification results** on our benchmark: judge whether a claim is factually true or false with external knowledge (Wikipedia or Web articles) as evidence.

Prompt	model	Edit-dis↓	WO↑	BS-F1↑	STS↑	Human
no-ques	ChatGPT	**0.207**	**0.864**	0.953	0.937	10
no-ques	GPT-4	0.275	0.789	0.954	0.931	28
with-ques	ChatGPT	0.222	0.853	**0.956**	**0.941**	13
with-ques	GPT-4	0.286	0.776	0.953	0.935	15

Table 6: **Revision evaluation** by intrinsic metrics and human (how many responses are preferred). Edit distance (**Edit-dis**) and word overlap (**WO**) between revised and the original responses. BERTScore (**BS-F1**) and semantic textual similarity (**STS**) based on SimCSE between the revised responses and human annotations.

Factcheck-GPT: implemented based on *langchain*.
SerpAPI retrieved evidence and GPT-4 served as the verifier.

Yuxia Wang et al. Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Take-Aways: Fact Checking of Machine-Generated Misinformation

- **Overall Findings**
  - Evidence retrieval significant bottleneck (only half of automatically retrieved evidence relevant to claim)
  - Factual inaccuracies difficult for LLMs to correct automatically (F1 of 0.63 for veracity prediction even with external knowledge)
  - Automatically evaluating the edited responses is difficult – intrinsic measures such as edit distance and semantic similarity are misaligned with human preferences

- **Future Possibilities**
  - Expand benchmark, including to more languages
  - Dealing with inter-claim dependencies
  - Better automatic judgement of relevance of retrieved evidence

Yuxia Wang et al. Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. 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
- **Post-Hoc Detection and Correction of Factual Errors**
  - Fact Checking and Correction of Machine-Generated Content
- **Probing the Parametric Knowledge of Language Models**
  - A Unified Framework for Input Feature Attribution Methods
  - Detecting Knowledge Conflicts of Language Models
- **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 the LM's parametric knowledge used to arrive at a LM's 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

- *Provides a human-interpretable explanation of the model's encoded parametric knowledge*

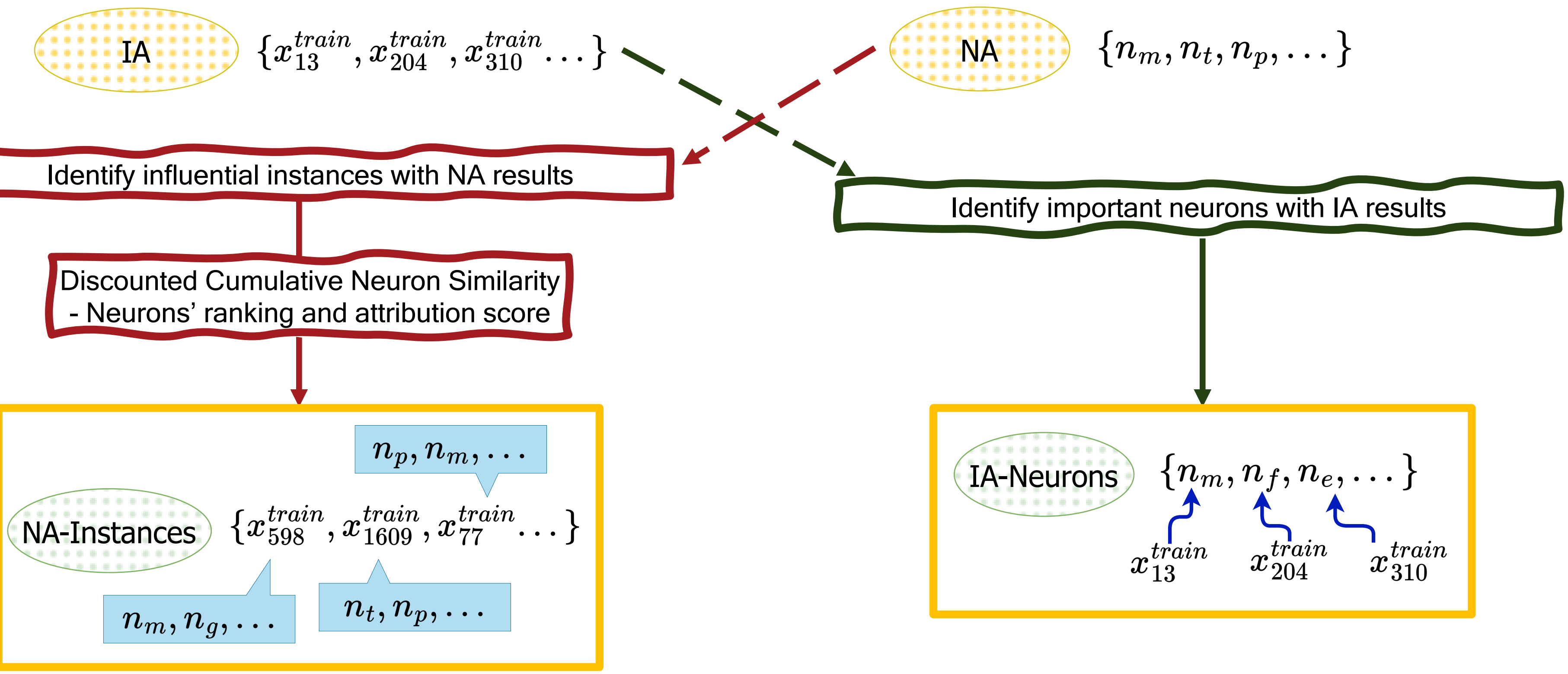
Neuron Attribution (NA) : Locates **specific neurons** that hold the most important parametric knowledge

- *Provides a 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

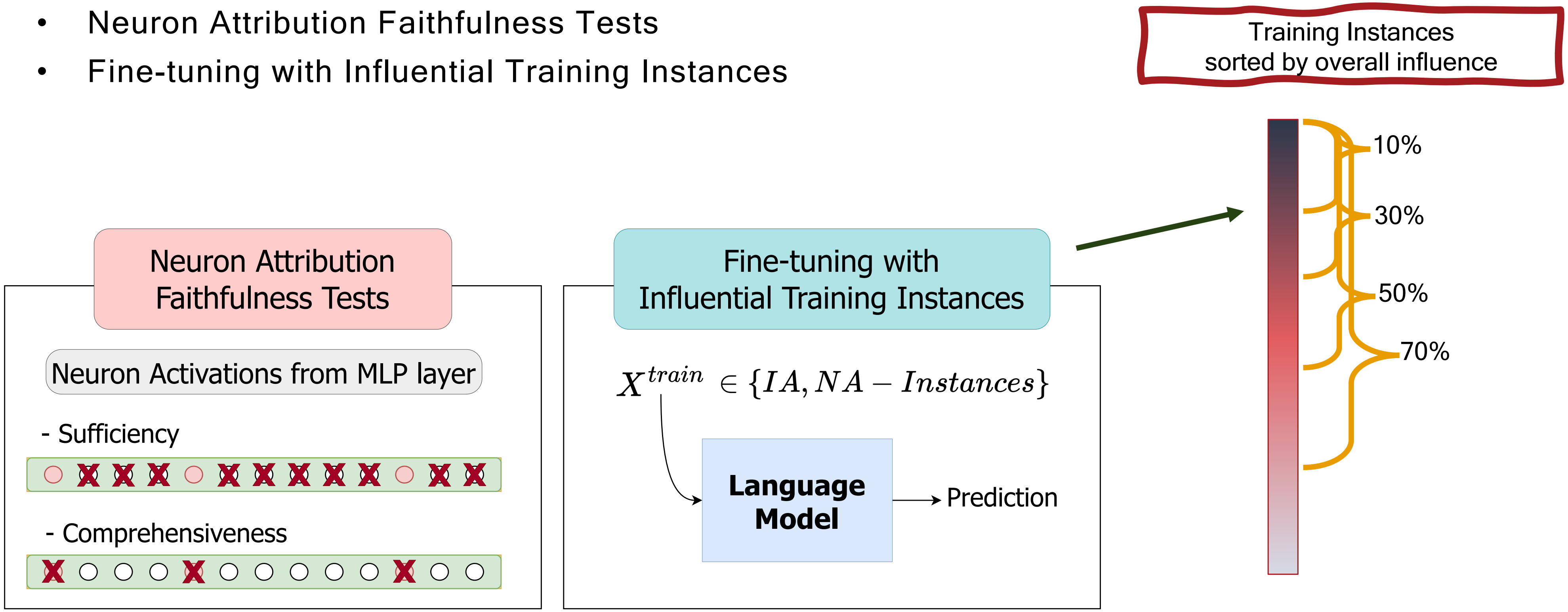
1) Aligning the Results of Attribution Methods



An Evaluation Framework for Attribution Methods

2) Tests

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



Experimental Set-up

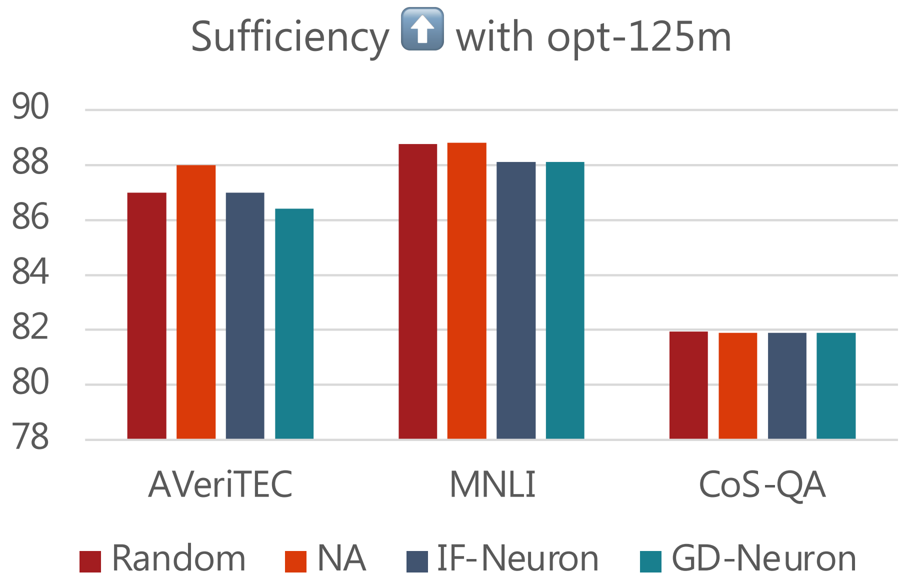
- Instance Attribution
  - Influence Function (IF) (Koh and Liang, 2017), Gradient Similarity (GS) (Charpiat et al., 2019)
- Neuron Attribution
  - The application of Integrated Gradient (Dai et al., 2022)

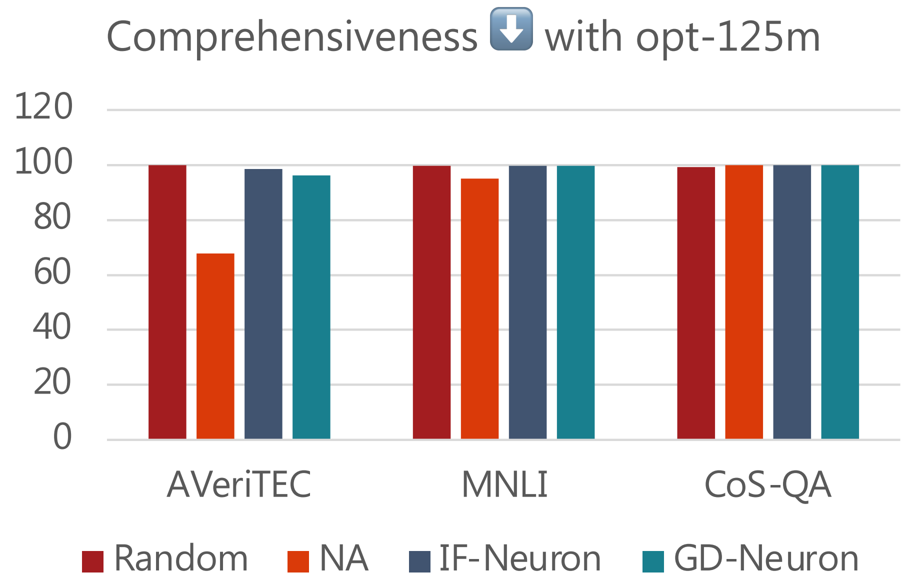
- Datasets
  - AVeriTeC (Fact-checking) / MNLI (Natural language inference) / Commonsense QA (Question Answering)

- 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





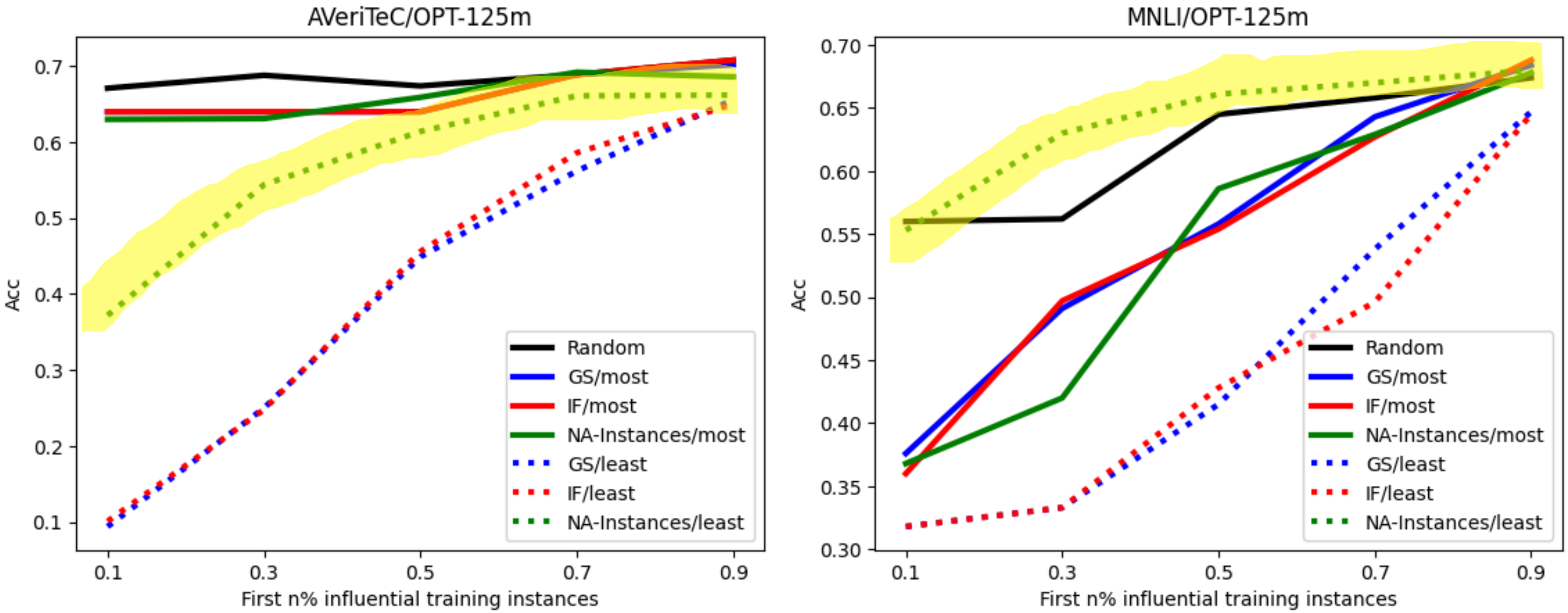
Evaluation metrics

- 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

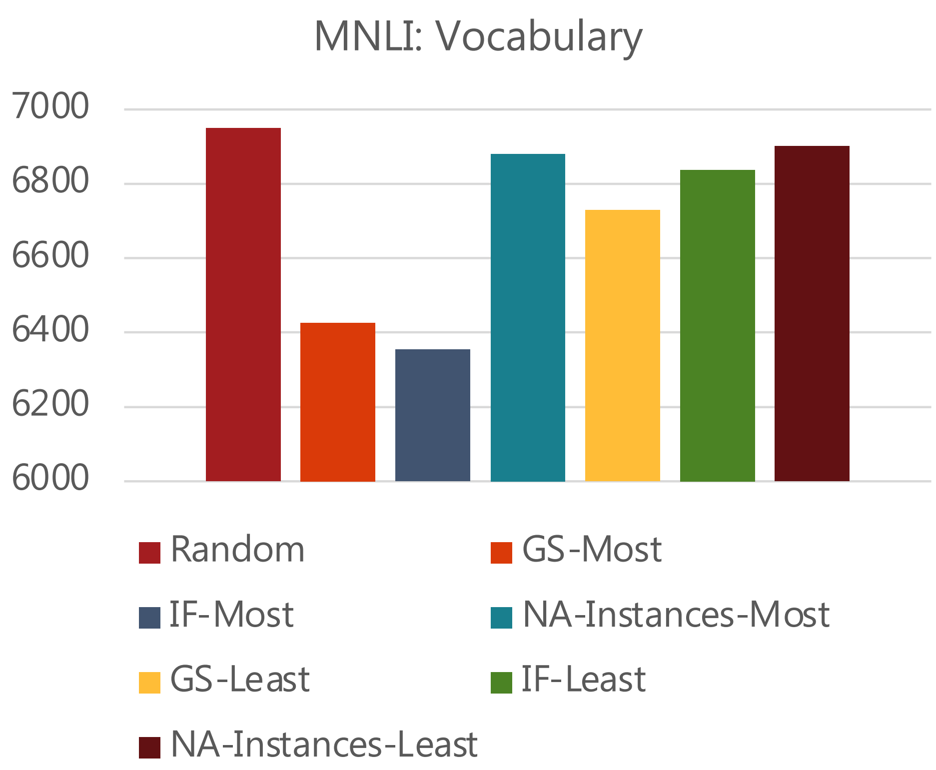
Fine-tuning with Influential Training Instances

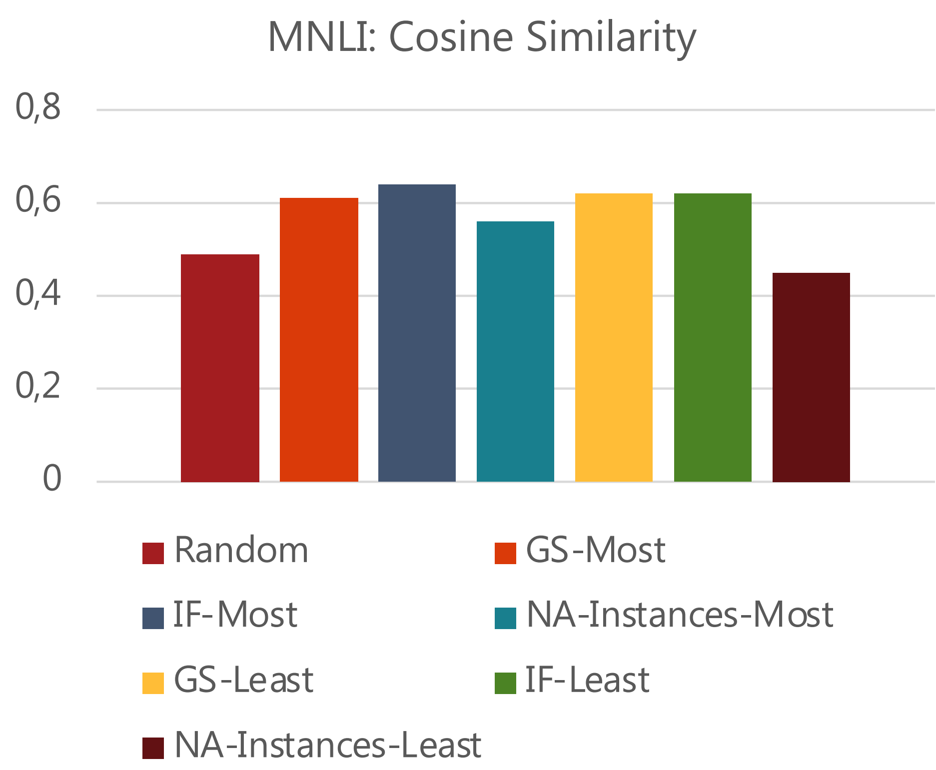


- NA-Instances-Least shows better performance than other least methods
- Counter-intuitive: why would IF-Least perform so well?

➢ Hypothesis: lack of diversity in selected instances

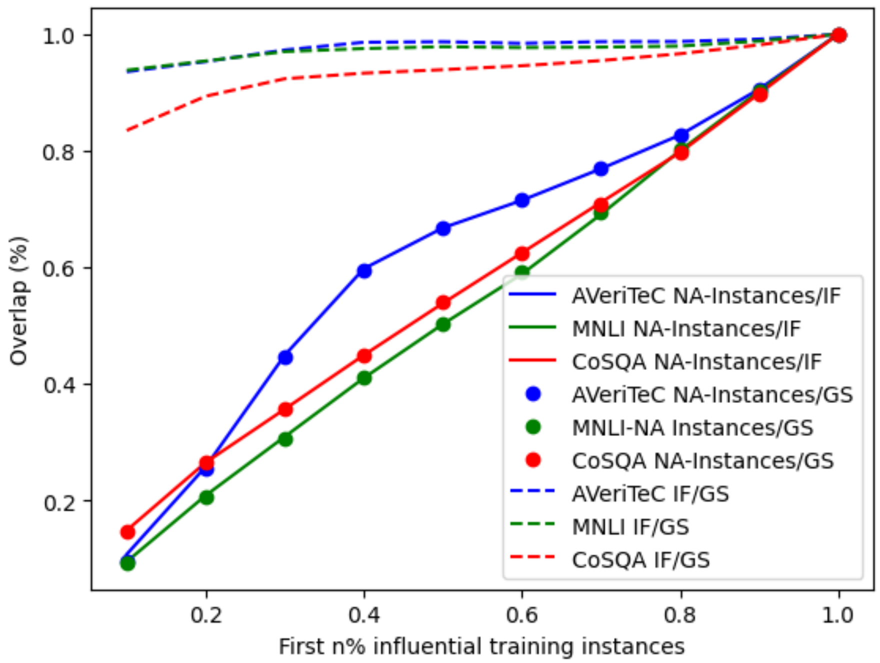
Diversity Analysis on the Group of Influential Training Instances





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

Overlap Analysis of Attribution Methods

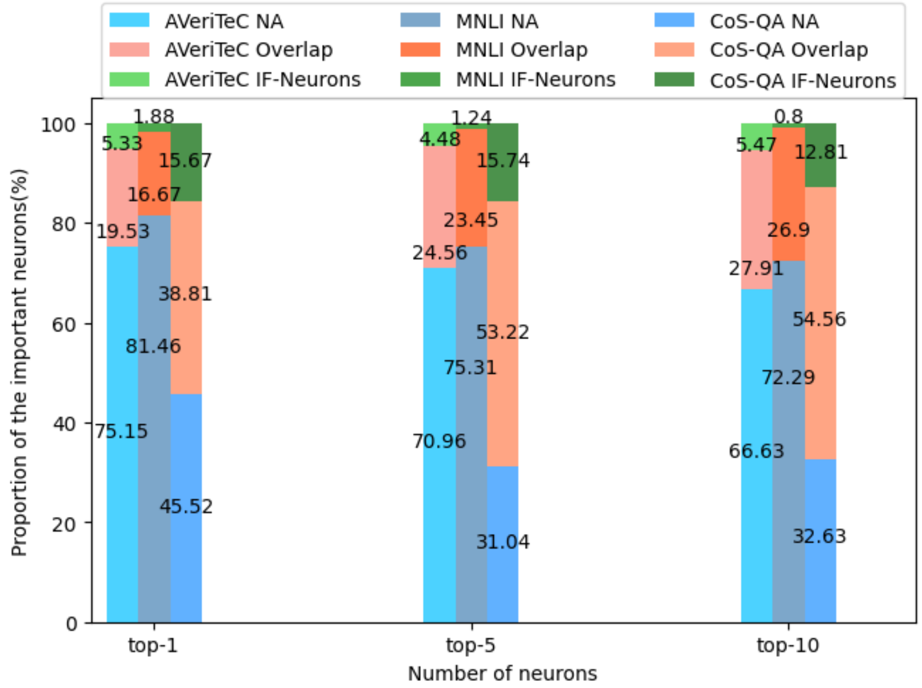


% of training instances at the intersection of the first n% influential instances discovered by a two of the attribution methods ∈ {IF, NA-Instances, and GS}

- 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

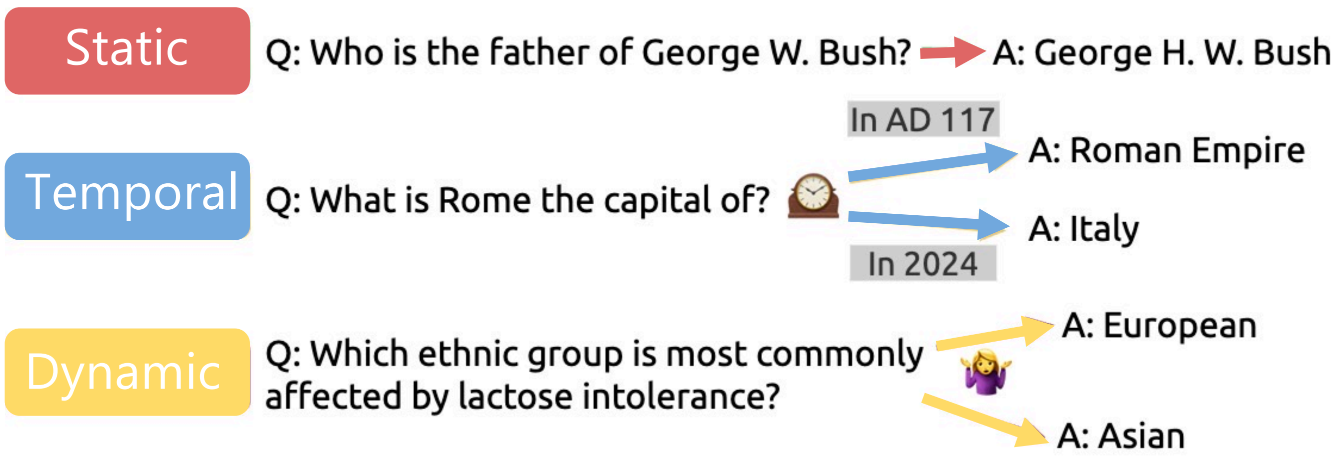
- We assess the sufficiency and comprehensiveness of the explanations for Instance Attribution and Neuron Attribution with different faithfulness tests
- We confirm that Instance Attribution and Neuron Attribution result in different explanations about the knowledge responsible for the test prediction
- The faithfulness tests suggest that the **neurons are not sufficient nor comprehensive enough** to fully explain the parametric knowledge used for the test prediction
- We hypothesise that this is due to the importance of the **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
- **Post-Hoc Detection and Correction of Factual Errors**
  - Fact Checking and Correction of Machine-Generated Content
- **Probing the Parametric Knowledge of Language Models**
  - A Unified Framework for Input Feature Attribution Methods
  - Detecting Knowledge Conflicts of Language Models
- **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

*

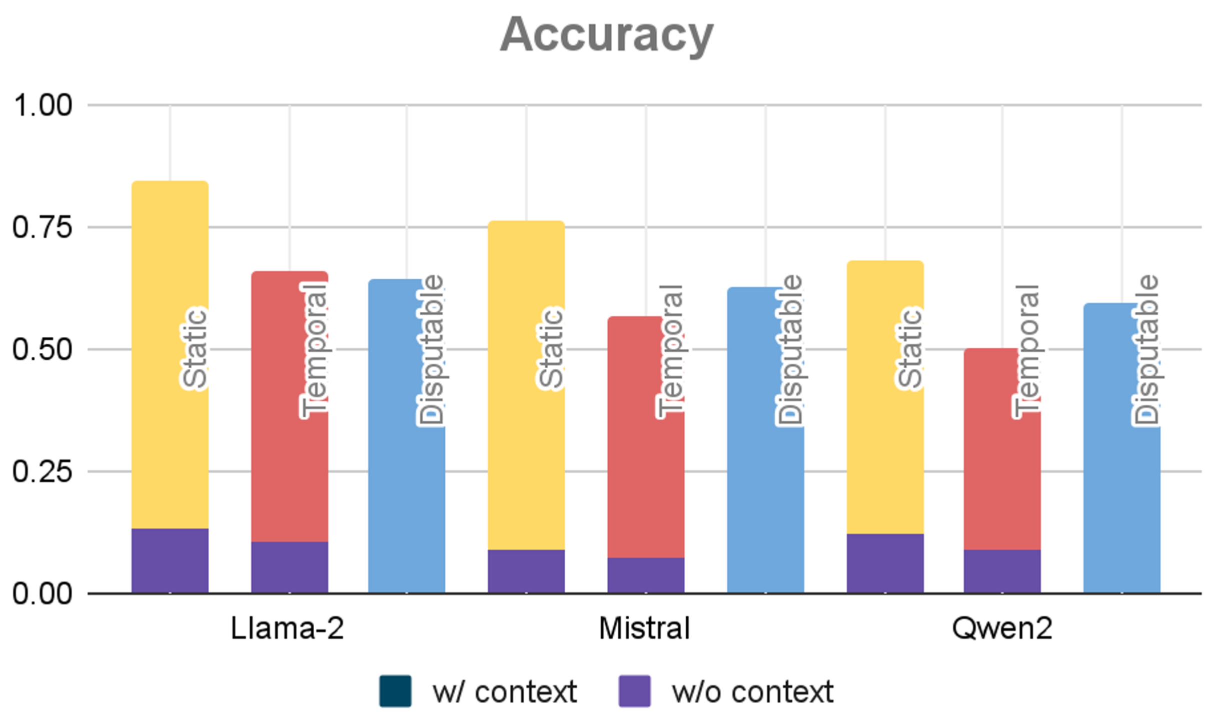
- Wikipedia:Controversial articles

0–9

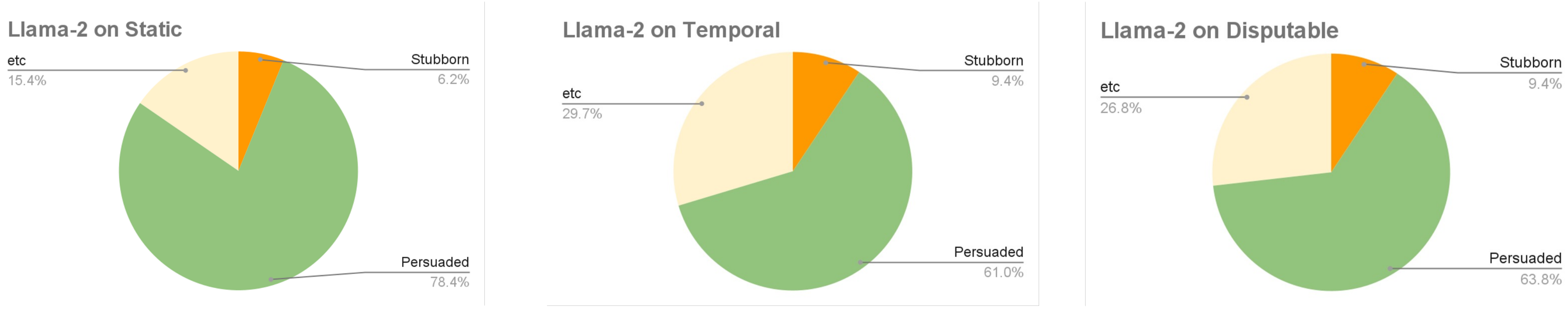
- Talk:2G spectrum case
- Talk:4B movement
- Talk:4chan
- Talk:4chan/Archive 16
- Talk:6ix9ine
- Talk:7 World Trade Center
- Talk:8chan
- Talk:9/11 conspiracy theories
- Talk:9/11 conspiracy theories regarding Jews or Israel
- Talk:10/40 window
- Talk:12 May Karachi riots
- Talk:40 Days for Life
- Talk:44M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

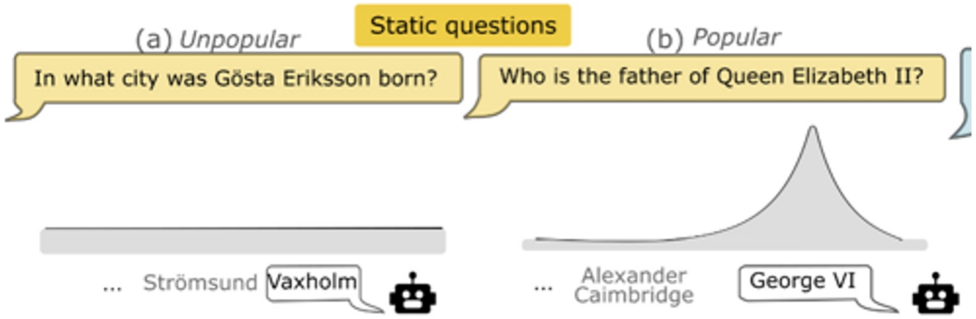


How do LMs perform on the dataset?

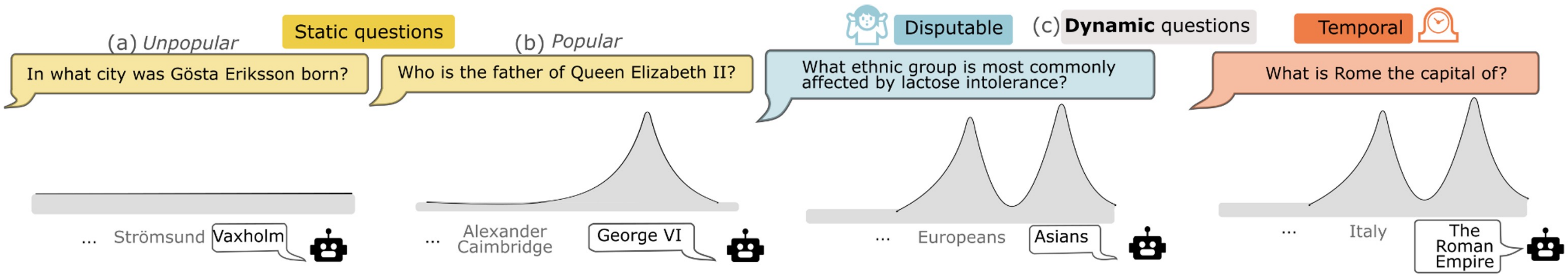


We see more **stubborn instances** in the dynamic partitions
-> Why are **dynamic** facts so **stubborn**?

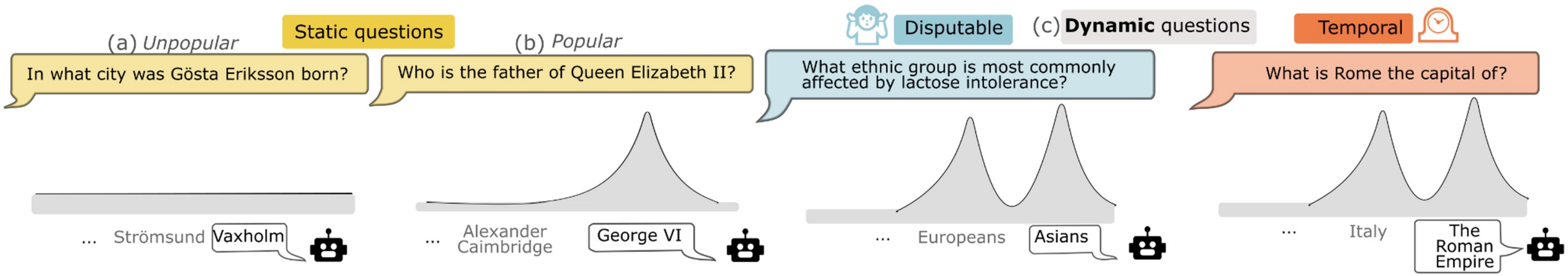
Intra-Memory Conflict in Output Distribution



Intra-Memory Conflict in Output Distribution



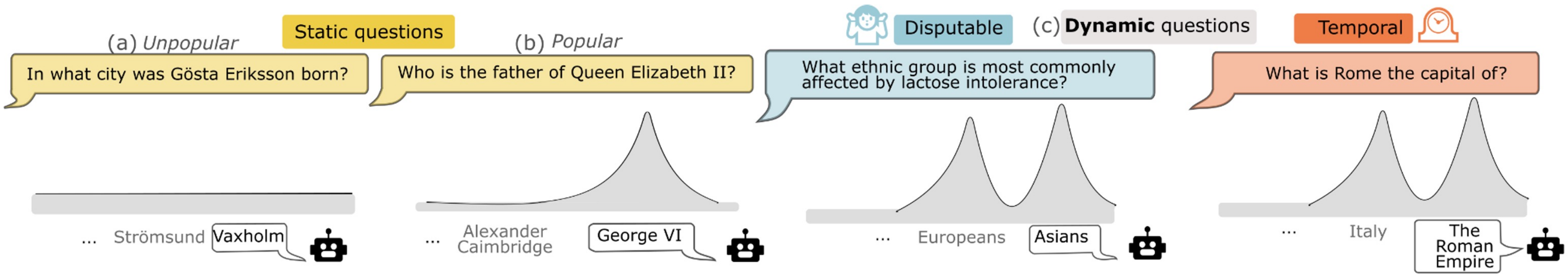
Intra-Memory Conflict in Output Distribution



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

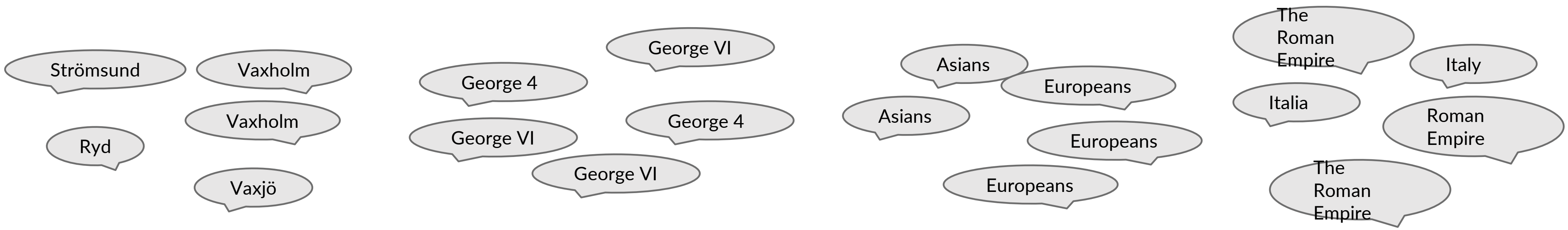
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*

Intra-Memory Conflict in Output Distribution

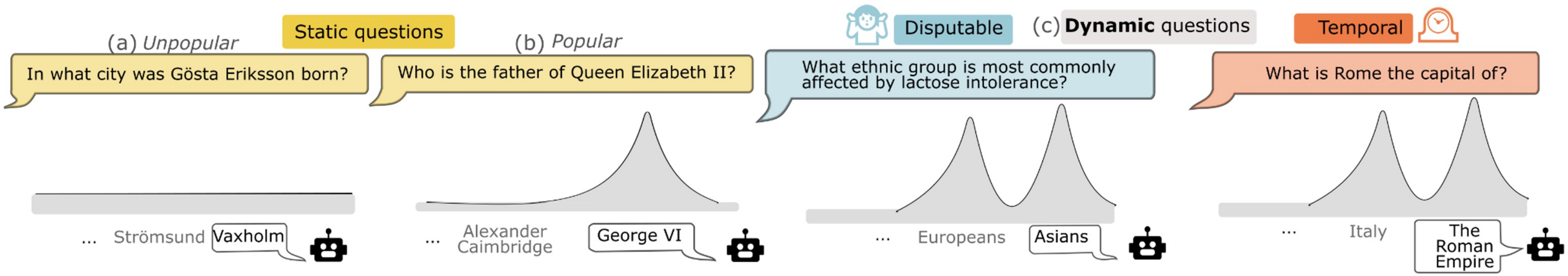


Dynamic facts should show greater _entropy_ across objects.

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

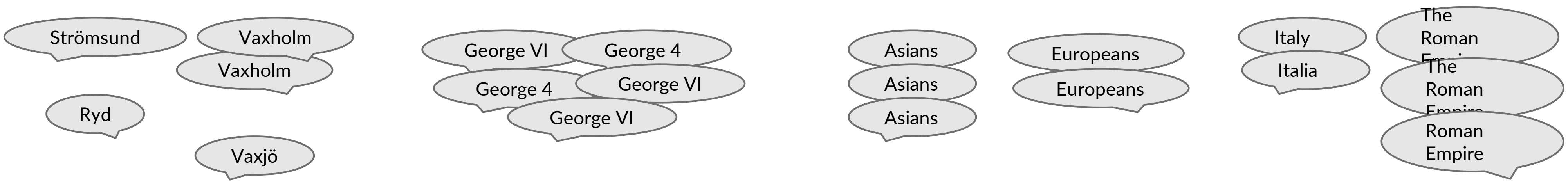


Intra-Memory Conflict in Output Distribution

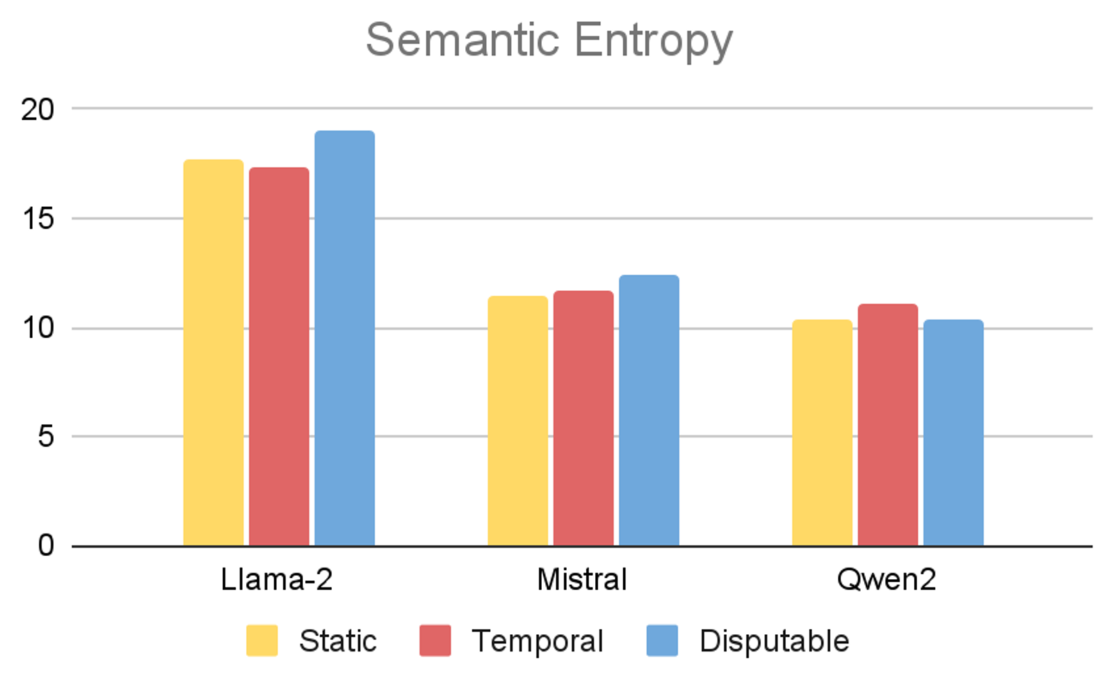


***Dynamic* facts should show greater *entropy* across objects.**

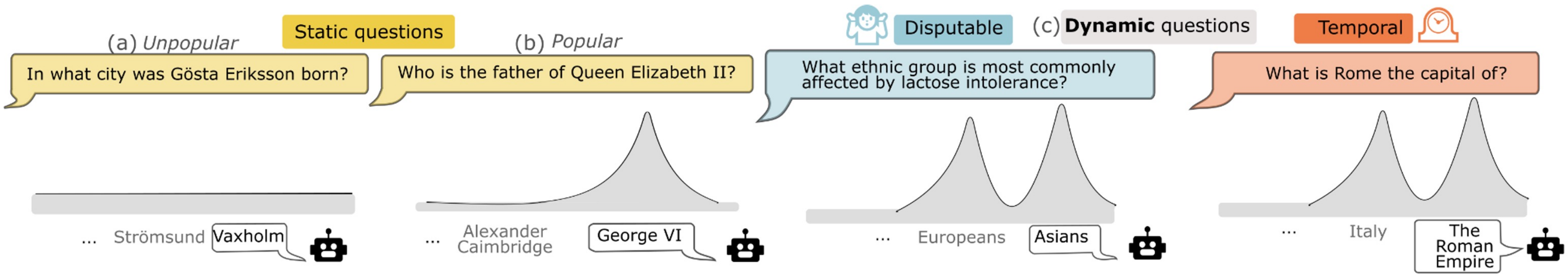
We evaluate this using *Semantic Entropy (Kuhn et al, 2023)*



However, this is not always the case

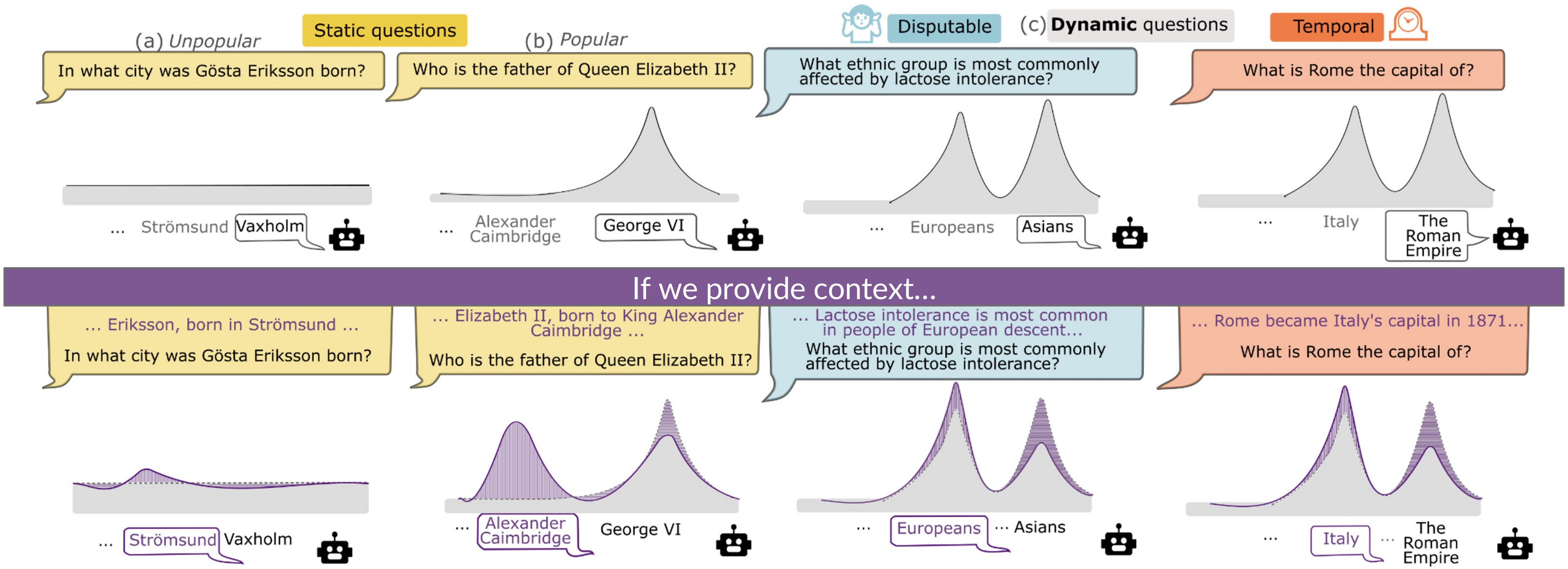


Intra-Memory Conflict

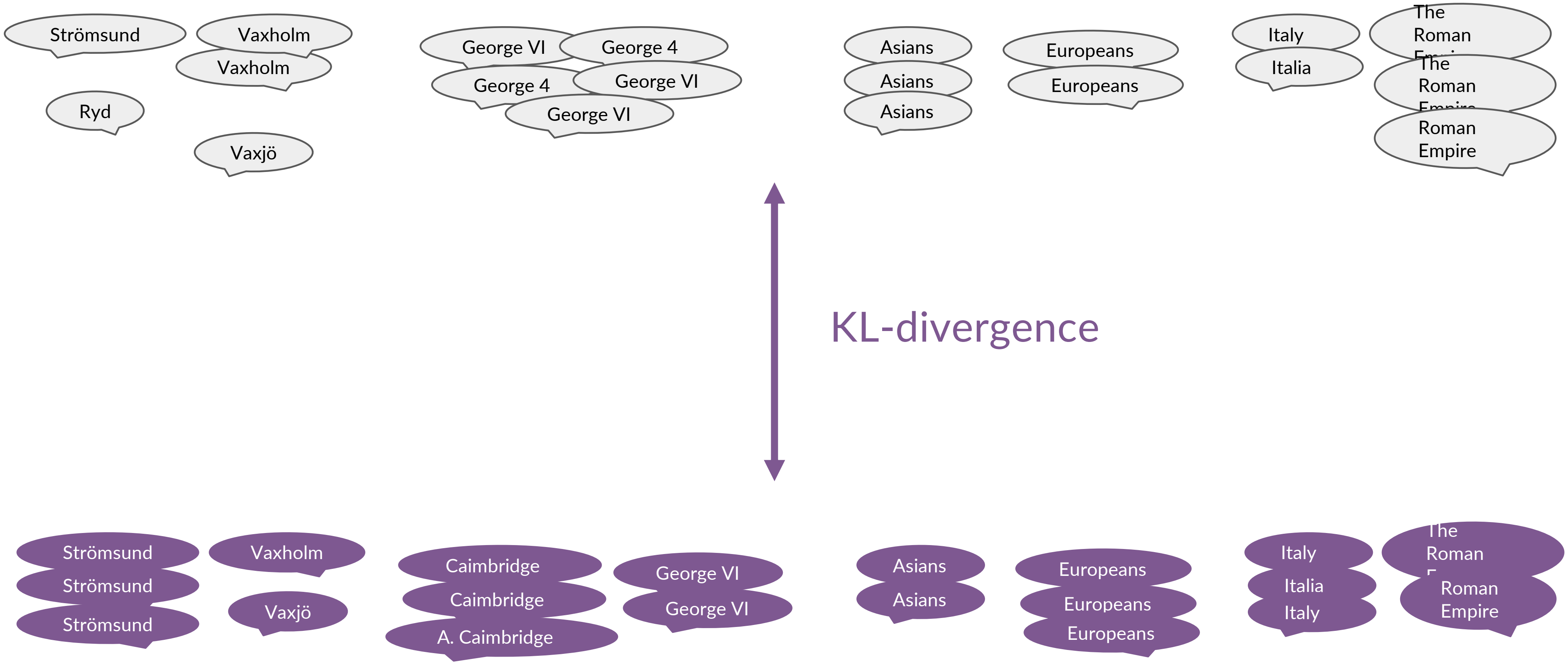


If we provide context…

Intra-Memory Conflict

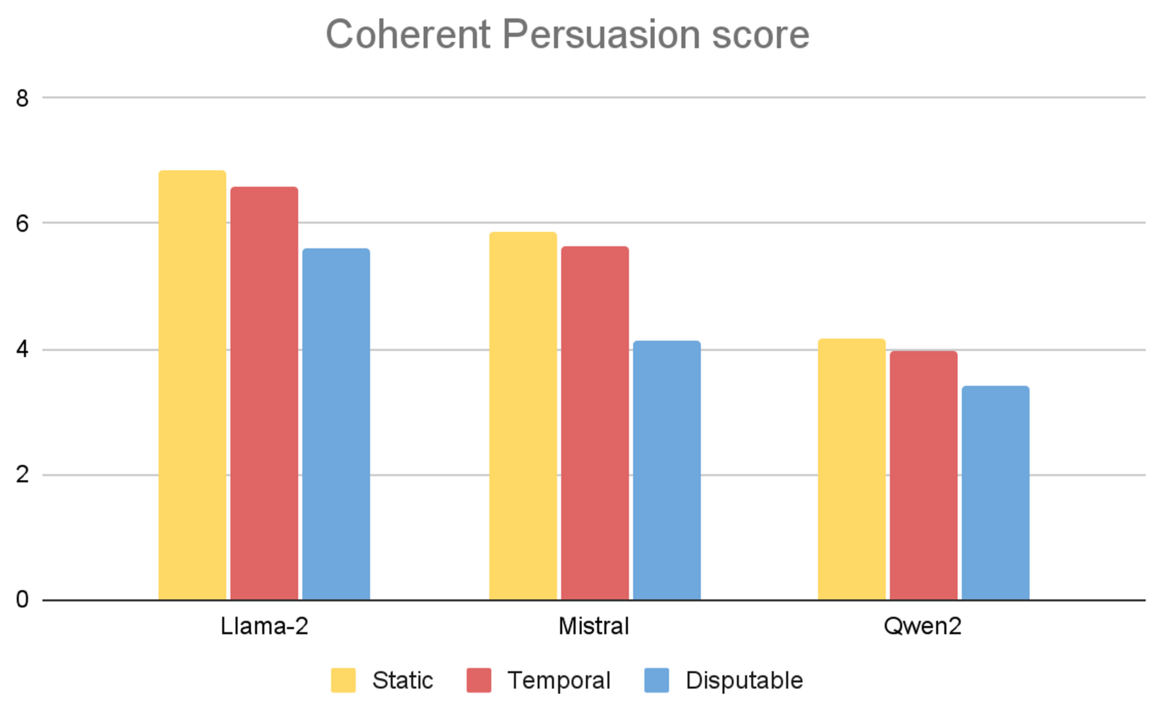


Coherent Persuasion Score



Persuasion Score across Partitions

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



Persuasion Score across Partitions

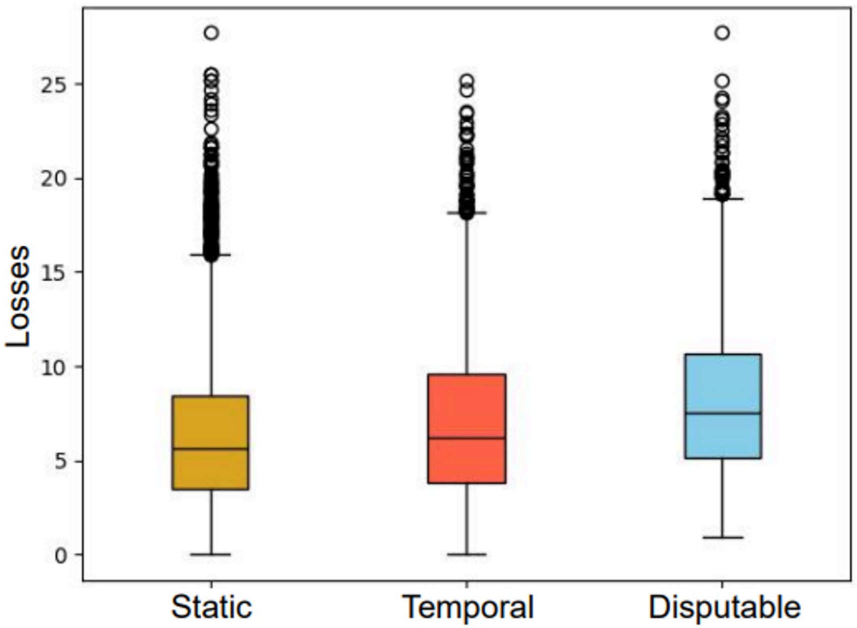
We see the **greatest persuasion score** for the **static dataset.**

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.

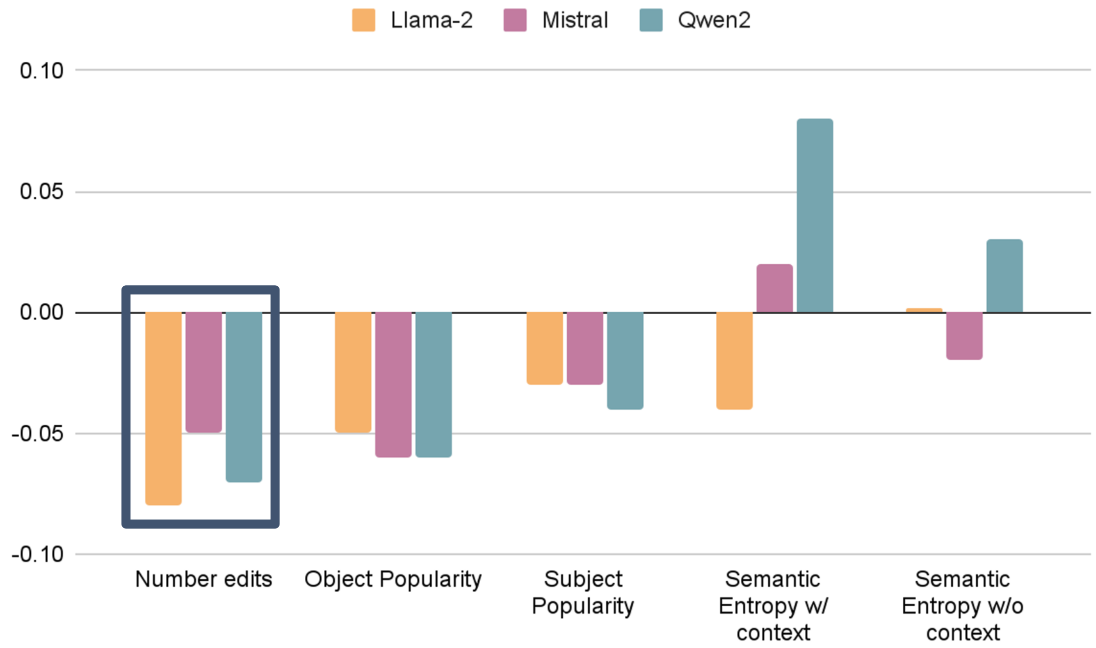
A higher loss indicates greater change required to steer the LM to output the target answer.

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.**

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, and greater 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
- **Post-Hoc Detection and Correction of Factual Errors**
  - Fact Checking and Correction of Machine-Generated Content
- **Probing the Parametric Knowledge of Language Models**
  - A Unified Framework for Input Feature Attribution Methods
  - Detecting Knowledge Conflicts of Language Models
- **Conclusion**
  - Wrap-Up and Outlook

Wrap-Up: Factuality Challenges of Large Language Models

- Despite seemingly high performance, LLMs suffer from **hallucinations**
- Potential to mislead public in novel ways
- Factuality challenges:
  - **Truthfulness**
  - **Unreliable evaluation**
  - Direct usage of misinformation
  - Lack of credible sourcing
  - Confident tone
  - Fluent style
  - Ease of access
  - Halo effect
  - Perceived as ”knowledge base”

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

Wrap-Up: Factuality Challenges of Large Language Models

- Threats posed by malicious LLM usage:
  - Personalised attacks
  - Style impersonation
  - Bypassing detection
  - Fake profiles

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

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

Thank you for your attention! Questions?

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.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, **Isabelle Augenstein**, Iryna Gurevych, Preslav Nakov. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 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.

Sara Vera Marjanović, **Isabelle Augenstein**, Christina Lioma. Investigating the Impact of Model Instability on Explanations and Uncertainty. In Findings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), August 2024.