Detecting Factual Errors of Large Language Models

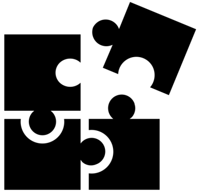
Isabelle Augenstein*

KnowLLM Workshop @ ACL
16 August 2024

*Partial credit for slides: Haeun Yu

UNIVERSITY OF COPENHAGEN

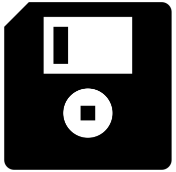
Factuality Challenges of Large Language Models



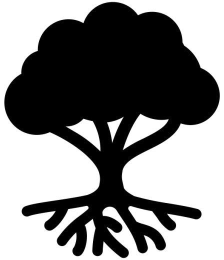
Citation Gaps

Truthfulness

Fluent Style



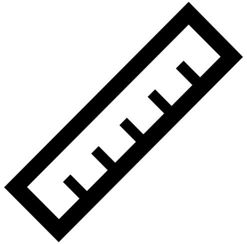
Outdated Knowledge



Grounding Deficiency

Confident Tone

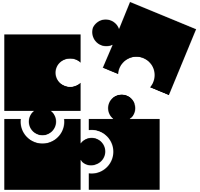
Halo Effect



Unreliable Evaluation

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

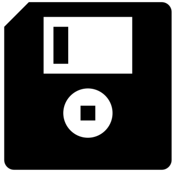
Factuality Challenges of Large Language Models



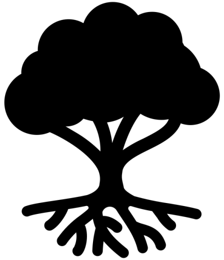
Citation Gaps

Truthfulness

Fluent Style



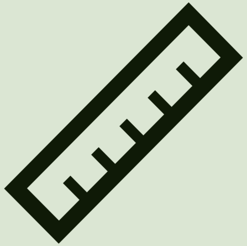
Outdated Knowledge



Grounding Deficiency

Confident Tone

Halo Effect



Unreliable Evaluation

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

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

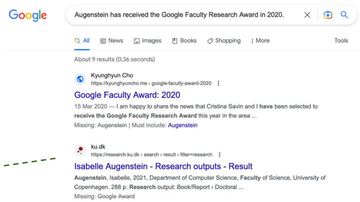
"Augenstein has published … and has received several awards, including the Google Faculty Research Award in 2020."

→ 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

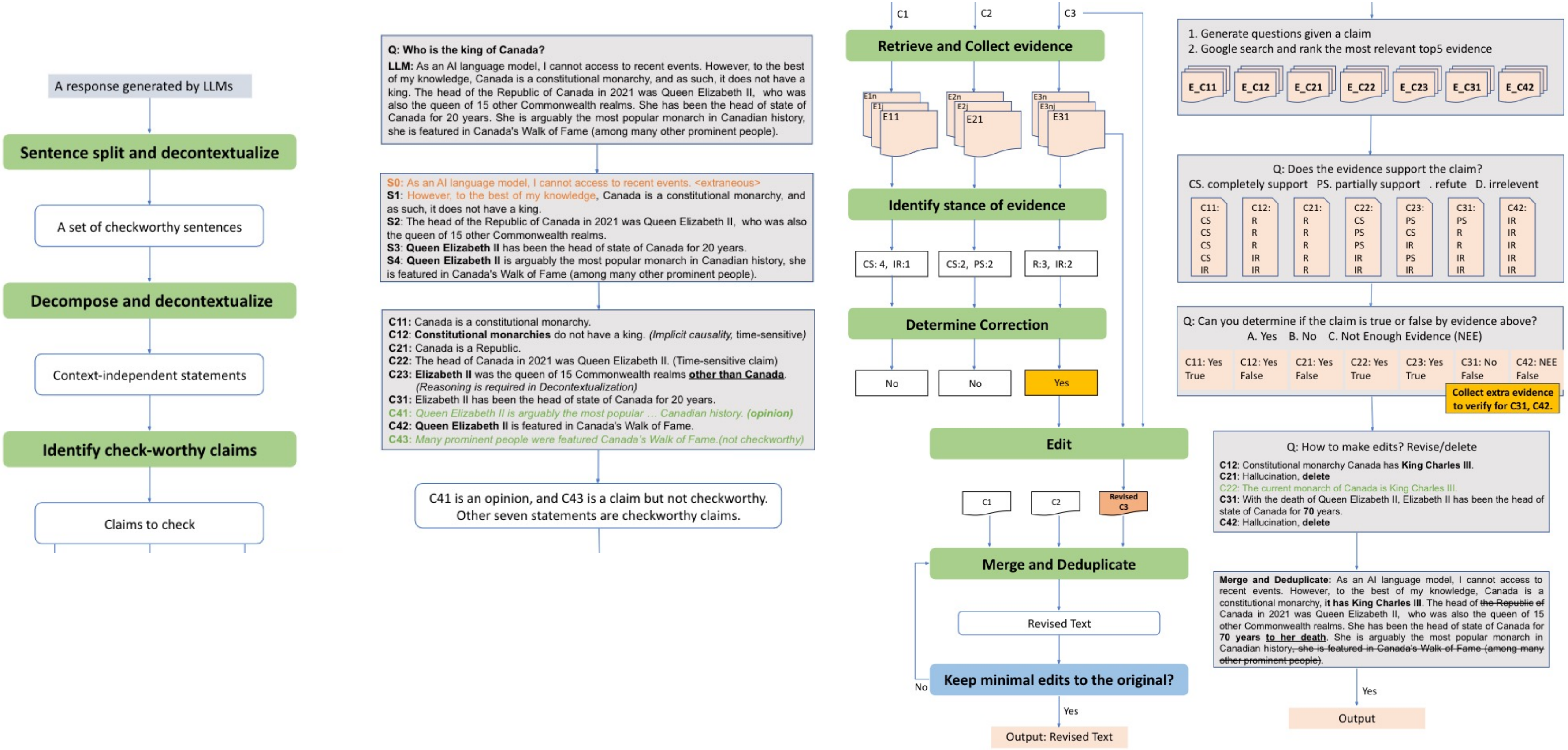
"Augenstein has published … and has received several awards, including the Google Faculty Research Award in 2020."

→ true

→ not enough info

→ false

Fact Checking and Correction of Machine-Generated Misinformation



Yuxia Wang et al. (2023). Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. CoRR, abs/2311.09000, November 2023.

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. (2023). Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. CoRR, abs/2311.09000, November 2023.

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*

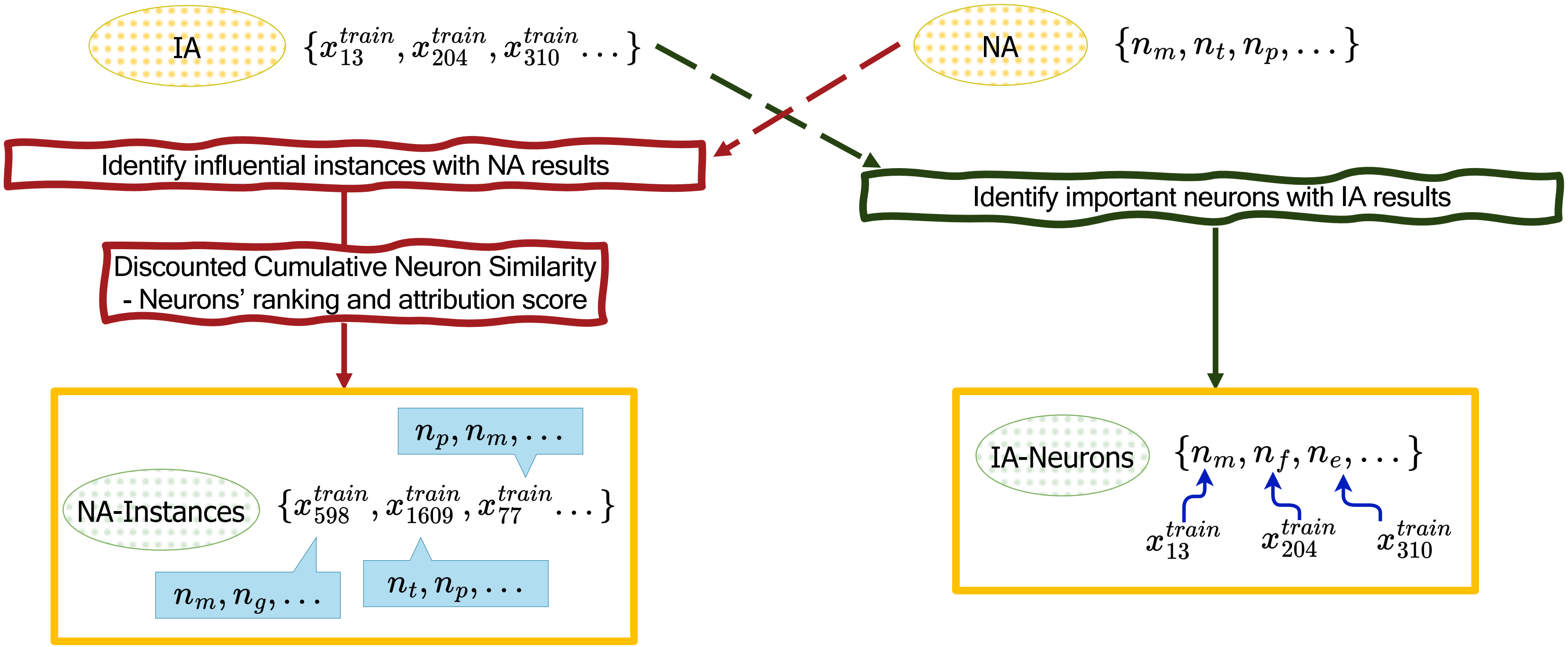
Neuron Attribution (NA) : Locates **specific neurons** that hold the most important parametric knowledge

- *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

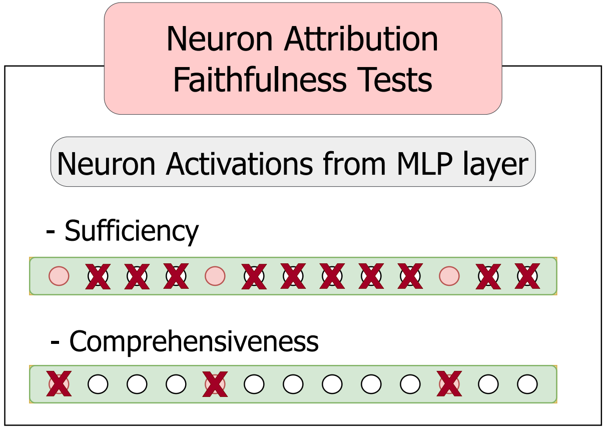
1) Aligning the Results of Attribution Methods

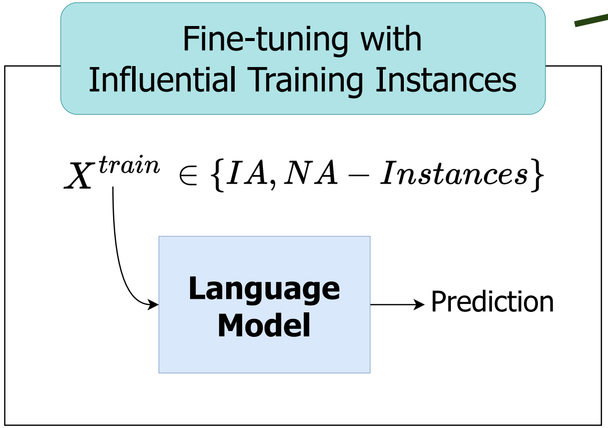


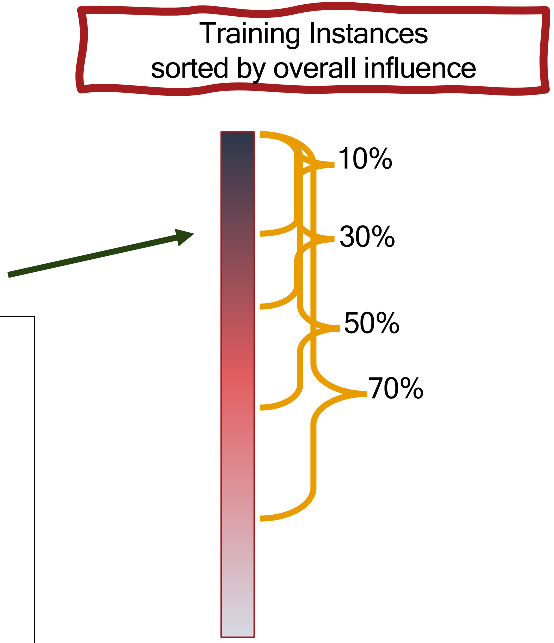
An Evaluation Framework for Attribution Methods

2) Tests

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







Experimental Set-up

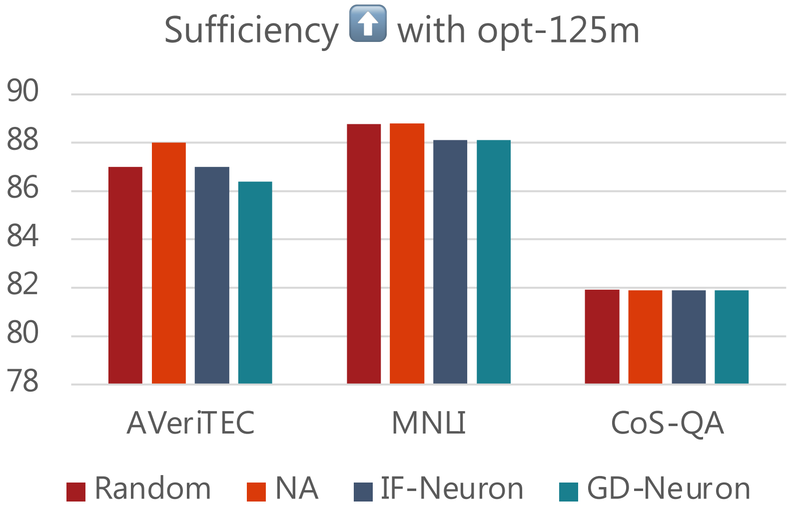
- Instance Attribution
  - Influence Function (IF) (Koh and Liang, 2017), Gradient Similarity (GS) (Charpiat et al., 2019)
- Neuron Attribution
  - The application of Integrated Gradient (Dai et al., 2022)

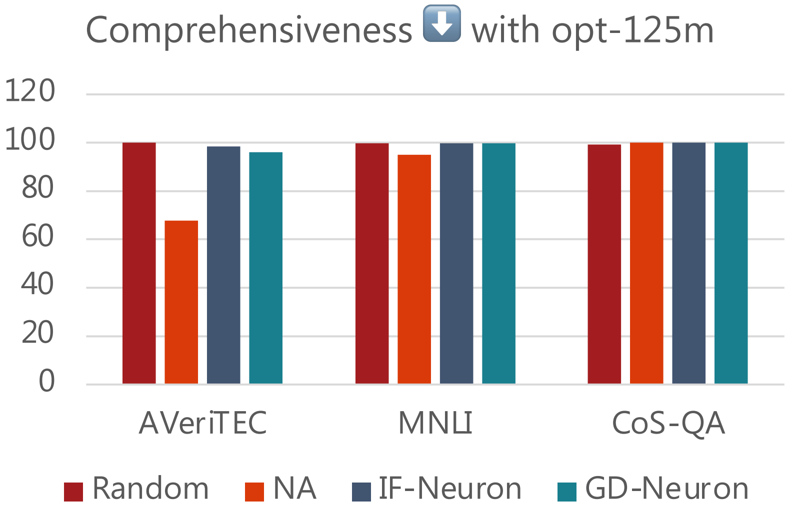
- Datasets
  - AVeriTeC (Fact-checking) / MNLI (Natural language inference) / Commonsense QA (Question Answering)

- 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

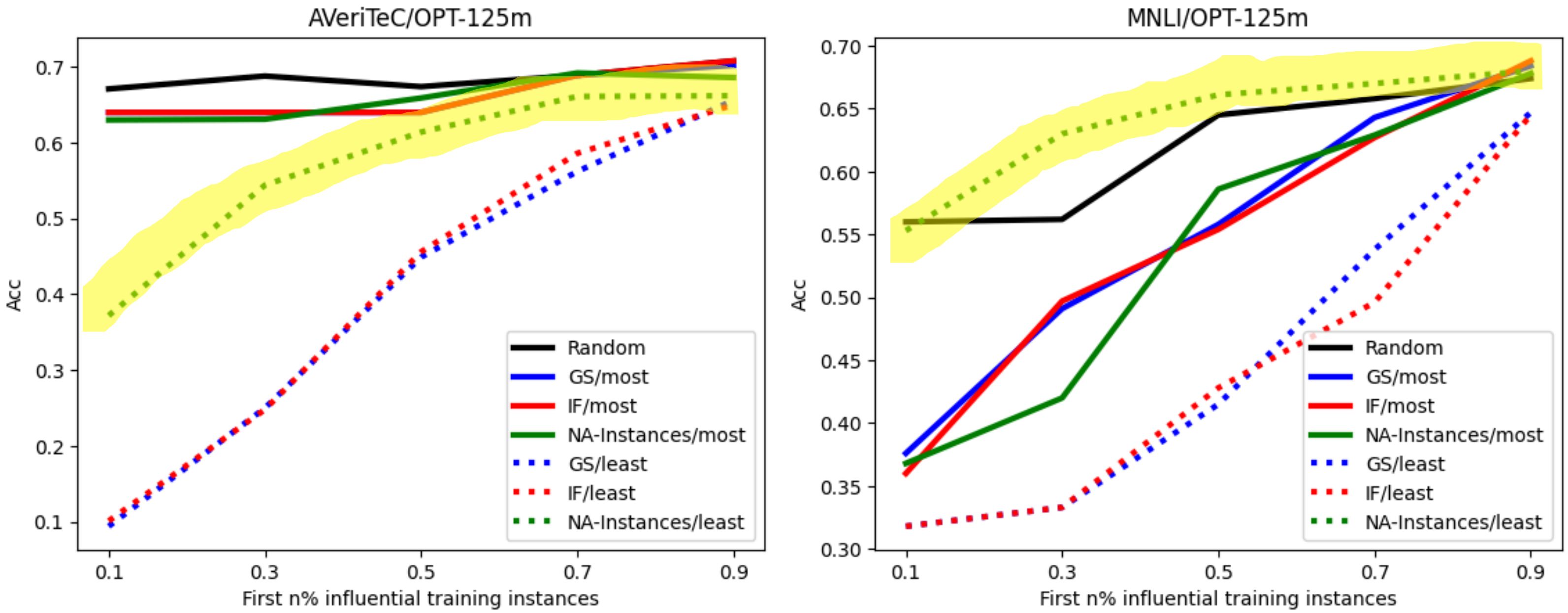
- Random: Randomly select the same number of neurons
- Sufficiency: Only use top-1 important neuron
- Comprehensiveness: Block top-100 neurons

Results

- Marginal differences among methods
- Only 1 neuron can recover prediction with above 70% accuracy

➢ Hypothesis: role of attention weights

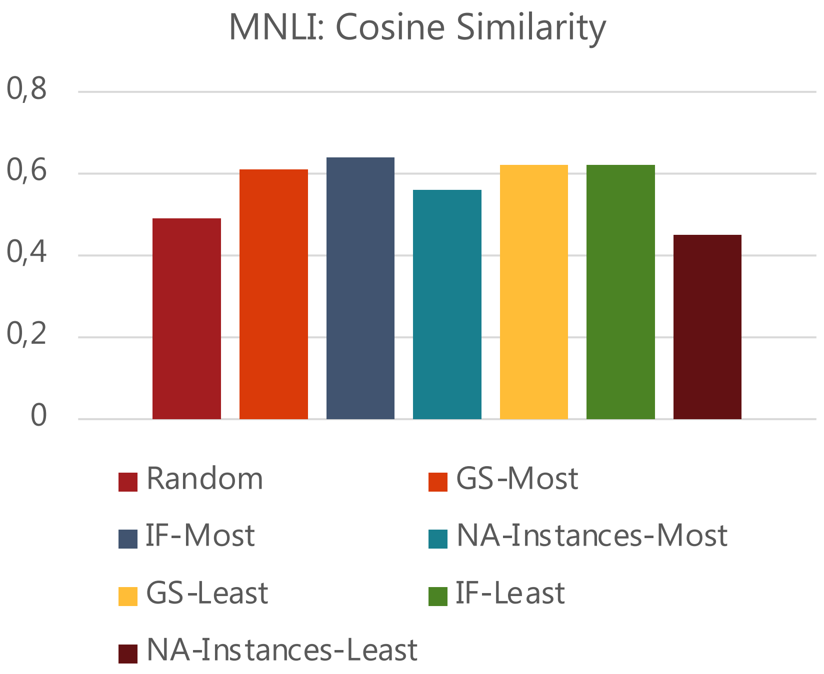
Fine-tuning with Influential Training Instances

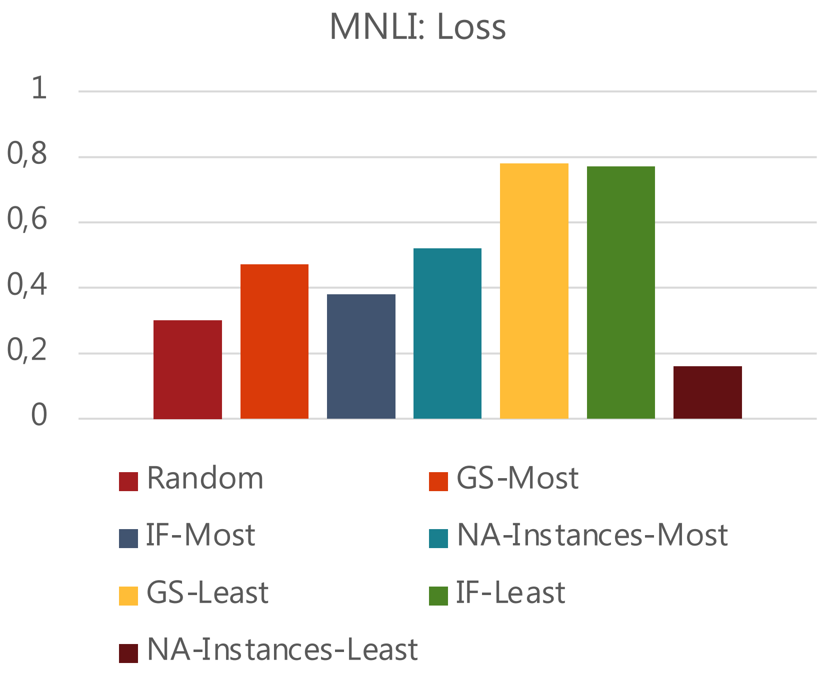


- NA-Instances-Least shows better performance than other least methods
- Counter-intuitive: why would IF-Least perform so well?

➢ Hypothesis: lack of diversity in selected instances

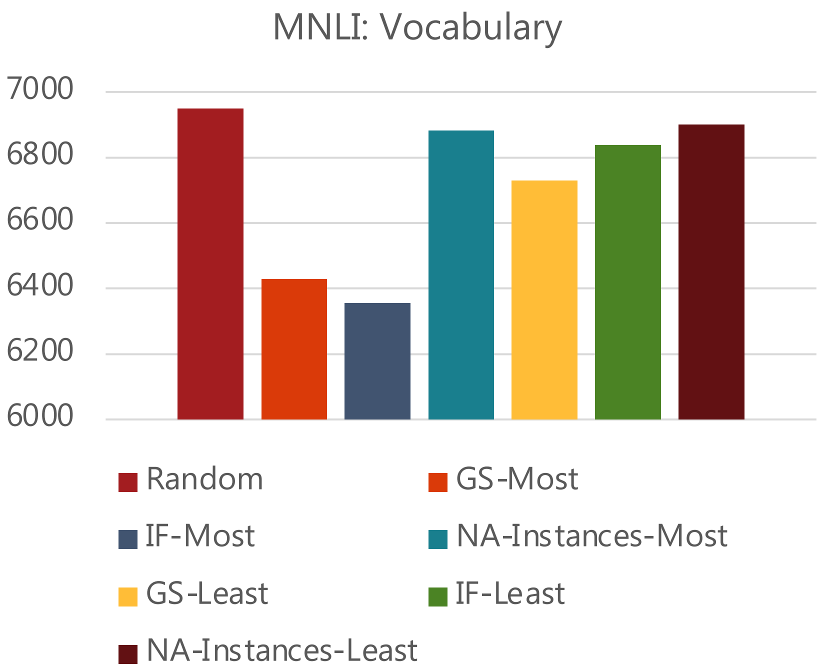
Diversity Analysis on the Group of Influential Training Instances

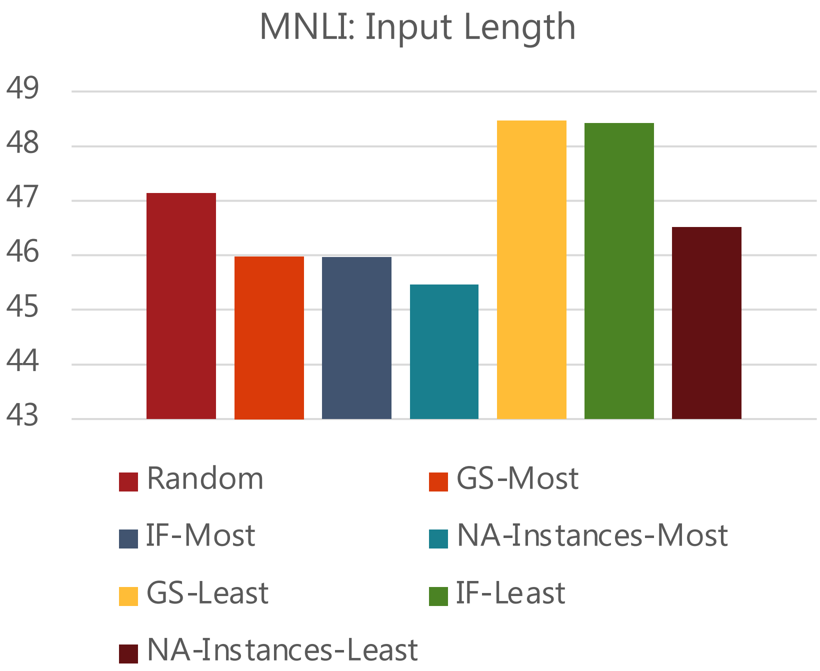




➢ NA-Instances-Least results in more diverse instances than Instance Attribution method GS

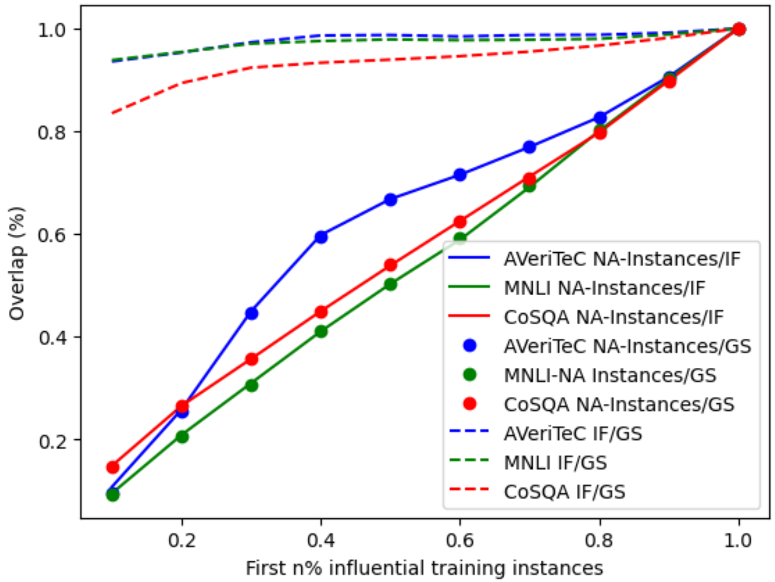
Diversity Analysis on the Group of Influential Training Instances





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

Overlap Analysis of Attribution Methods

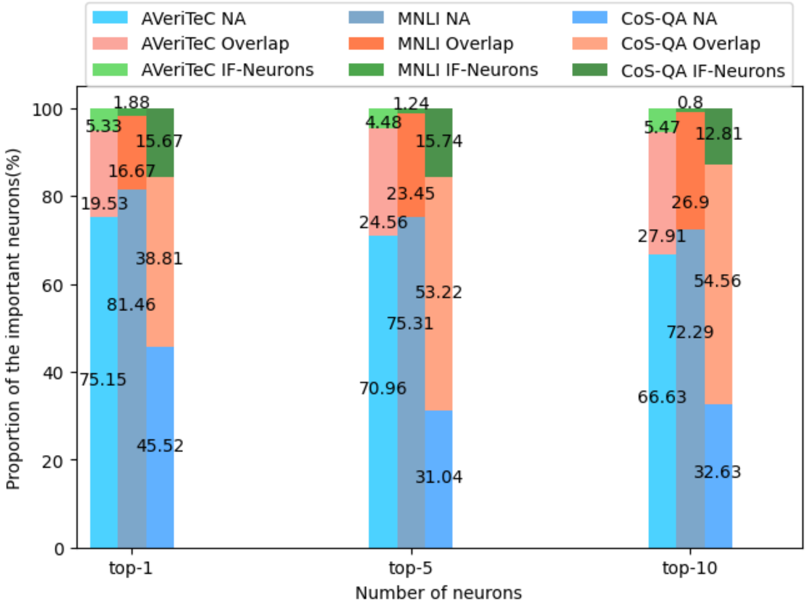


% of training instances at the intersection of the first n% influential instances discovered by a two of the attribution methods ∈ {IF, NA-Instances, and GS}

- 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
- 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

Knowledge Conflict and Fact Dynamicity

- 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

- Fact Dynamicity
  - Temporality: Facts that change over time
  - Disputability: Facts that vary depending on the point of view

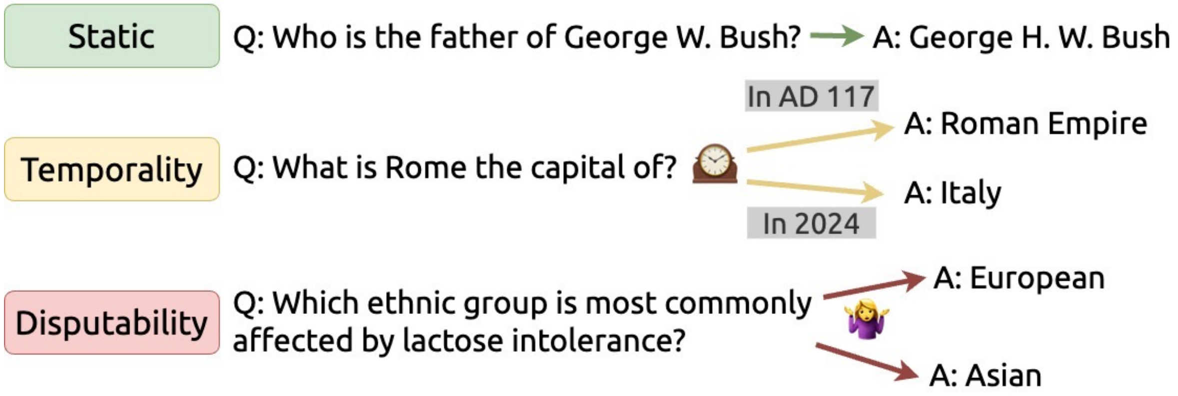
We investigate the interaction between intra-memory conflict and context-memory conflicts, using multiple natural causes of intra-memory conflict (i.e. fact ‘dynamicity’).

Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. From Internal Conflict to Contextual Adaptation of Language Models. CoRR, abs/2407.17023, July 2024.

DynamicQA

- Consists of 11,288 context-question pairs
- Featuring two different contexts and answers for the same question
- Based on Wikidata / Wikipedia edit history

	# of Questions	# of Instances
Static	2500	5000
Temporal	2495	4900
Disputable	694	1388



DynamicQA

Static / Temporal

- Based on PopQA (Wikidata based QA dataset)
- Given questions, identify context

- Identify temporal QA pair and static pair
  - If # edits > 1, temporal
  - Else, static

- For contexts, find the sentence from the Wikipedia article that mentions the object

Disputable

- Based on Wikipedia's list of controversial articles
- Given context, generate questions

- Identify reverted edits in Wikipedia edit logs

With two versions of Wikipedia edit history:

- Identify reverted word with edit distance
- Filter vandalism / synonym / paraphrasing
- Generate question with LM

DynamicQA

- Introducing a novel dataset of knowledge conflicts in the real world
  - Approximation of the degree of the knowledge conflict in the real-world
  - Statisticity: Number of monthly Wikipedia article views
  - Temporality: Number of Wikidata edits of object given same subject and relation
  - Disputability: The occurrence of the pair of reverted edit logs

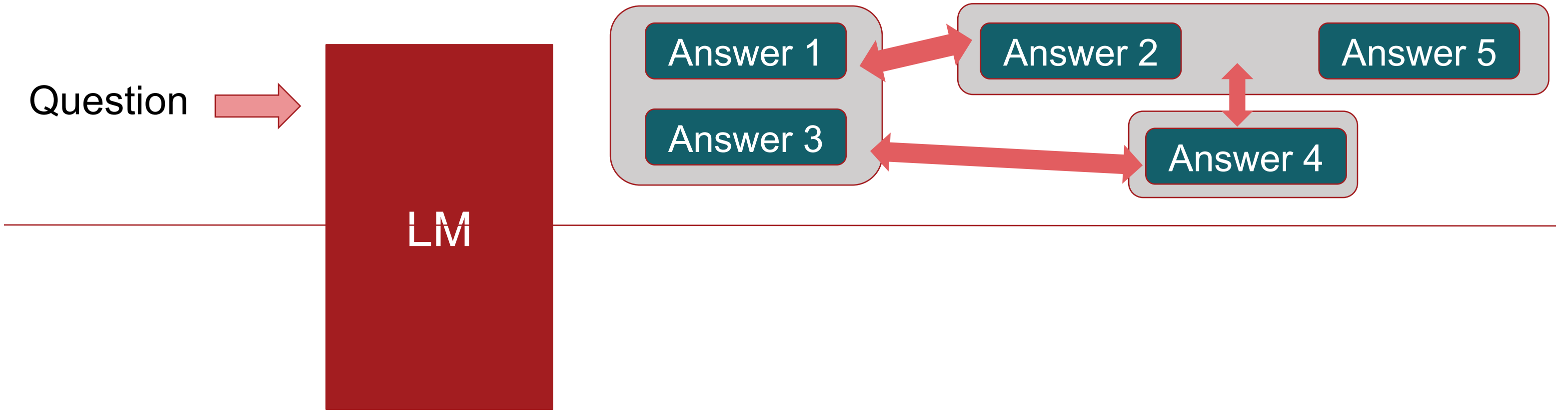
- Human Annotation on Disputable facts
  - Two annotators annotated each datapoint, and conflicts were resolved by the third annotator (Krippendorf's alpha of 0.44)

Measuring Intra-Memory Knowledge Conflict

1. Generate multiple answers using sampling
2. Group the answers by their semantic similarity -> Semantic sets (with NLI model)

Semantic Uncertainty (Kuhn et al., 2023) for the **Intra-Memory Conflict**

=> Entropy between the semantic sets

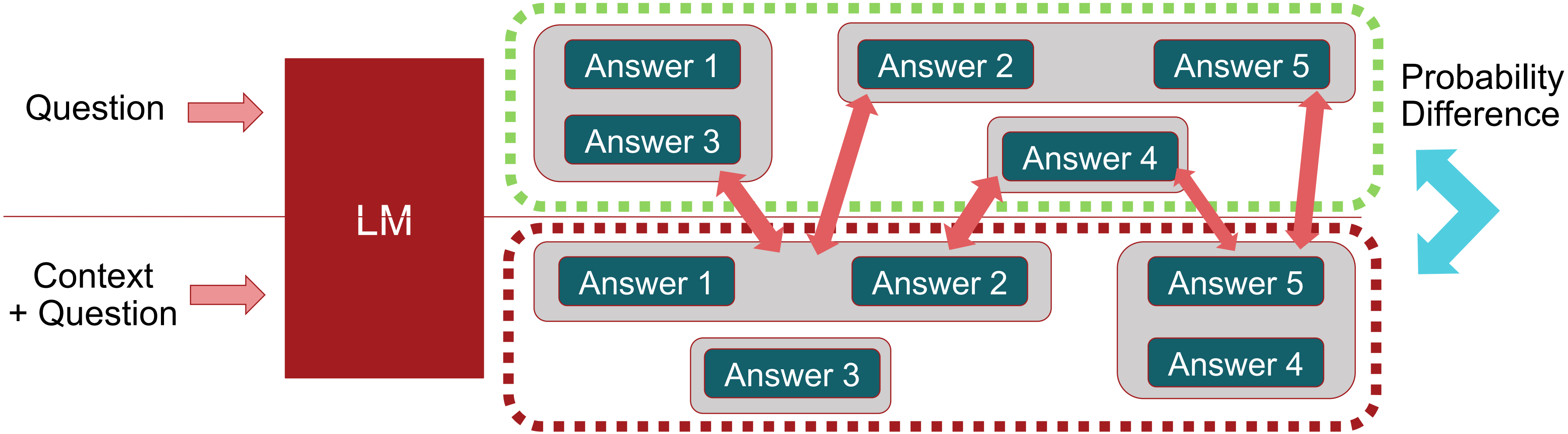


Measuring Context-Memory Knowledge Conflict

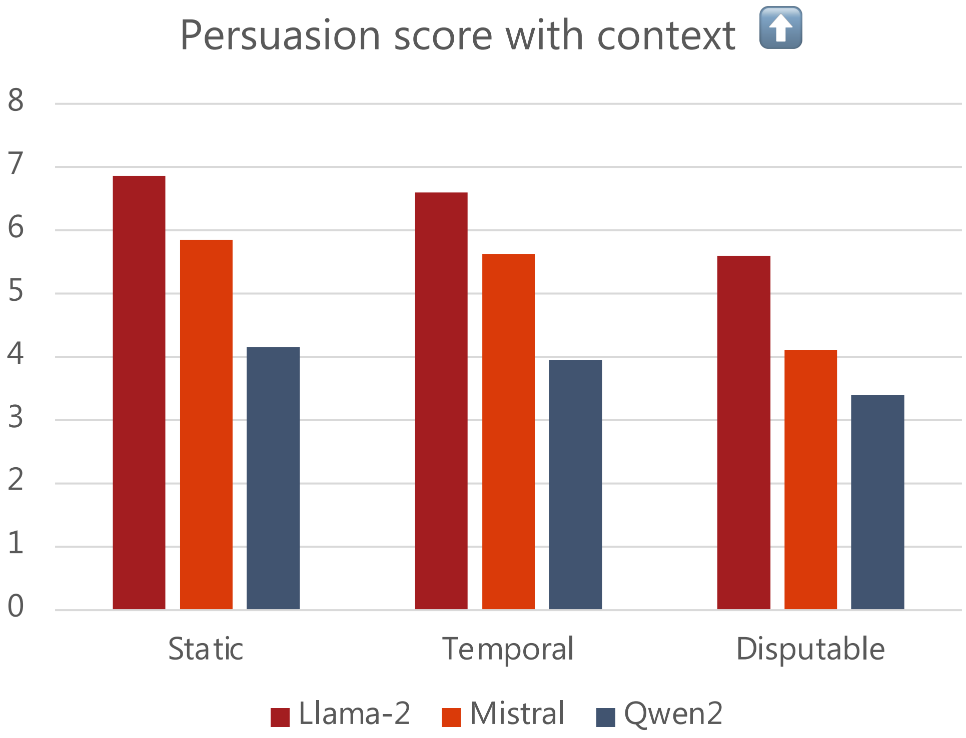
Coherent Persuasion score for the **Context-Memory Conflict**

Considers all possible answers from a LM

- Averaging the difference of probability distribution between all permutations of semantic sets from question and context+question

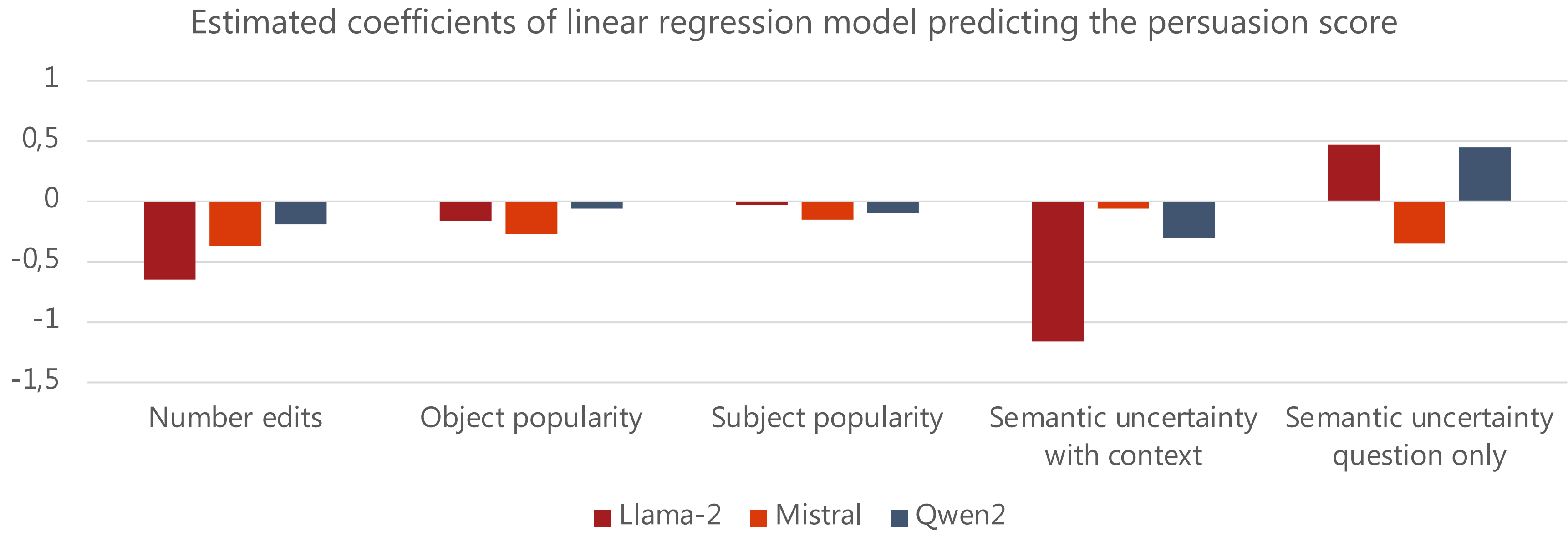


Are models more likely to change their predictions for dynamic facts?



- Unexpected Finding: Models **more easily persuaded to change predictions for static facts**
    - Those are expected to have smaller variability in the training dataset, and thus smaller intra-memory conflict
- Potential implications for **efficacy of retrieval-augmented generation**
    - Most commonly updated facts are the most difficult to adapt in the model

What are predictors of persuasion?



- Number of edits consistent strong inverse predictor for persuasion score
- Subject/object popularity insignificant effect
- Uncertainty of question with/without context not reliable predictor

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**. From Internal Conflict to Contextual Adaptation of Language Models. CoRR, abs/2407.17023, July 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, July 2024, to appear.

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, July 2024, to appear.

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, July 2024, to appear.

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-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. CoRR, abs/2311.09000, November 2023.

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, to appear.

Sara Vera Marjanović*, Haeun Yu*, Pepa Atanasova, Maria Maistro, Christina Lioma, **Isabelle Augenstein**. From Internal Conflict to Contextual Adaptation of Language Models. CoRR, abs/2407.17023, July 2024.