UNIVERSITEIT VAN AMSTERDAM

INDE lab

Bilateral Factuality Evaluation using Large Language Models

Bradley P. Allen

2025-02-04

Overview

- Improving the reliability of factuality evaluation by large language models (LLMs) is crucial for trustworthy AI
- We hypothesize that LLMs can better evaluate the factuality of statements by both verifying and refuting claims (a *bilateral* approach) versus only making a single true/false judgment (a *unilateral* approach)
- We tested both approaches using zero-shot chain-of-thought (CoT) prompting of a variety of LLMs against a benchmark comprised of 1,000 factual questions with true or false answers
- Joint work with Paul Groth (UvA), Filip Ilievski (VU), Thomas Ferguson (RPI), and Prateek Chhikara (USC/Mem0)

Factuality evaluation

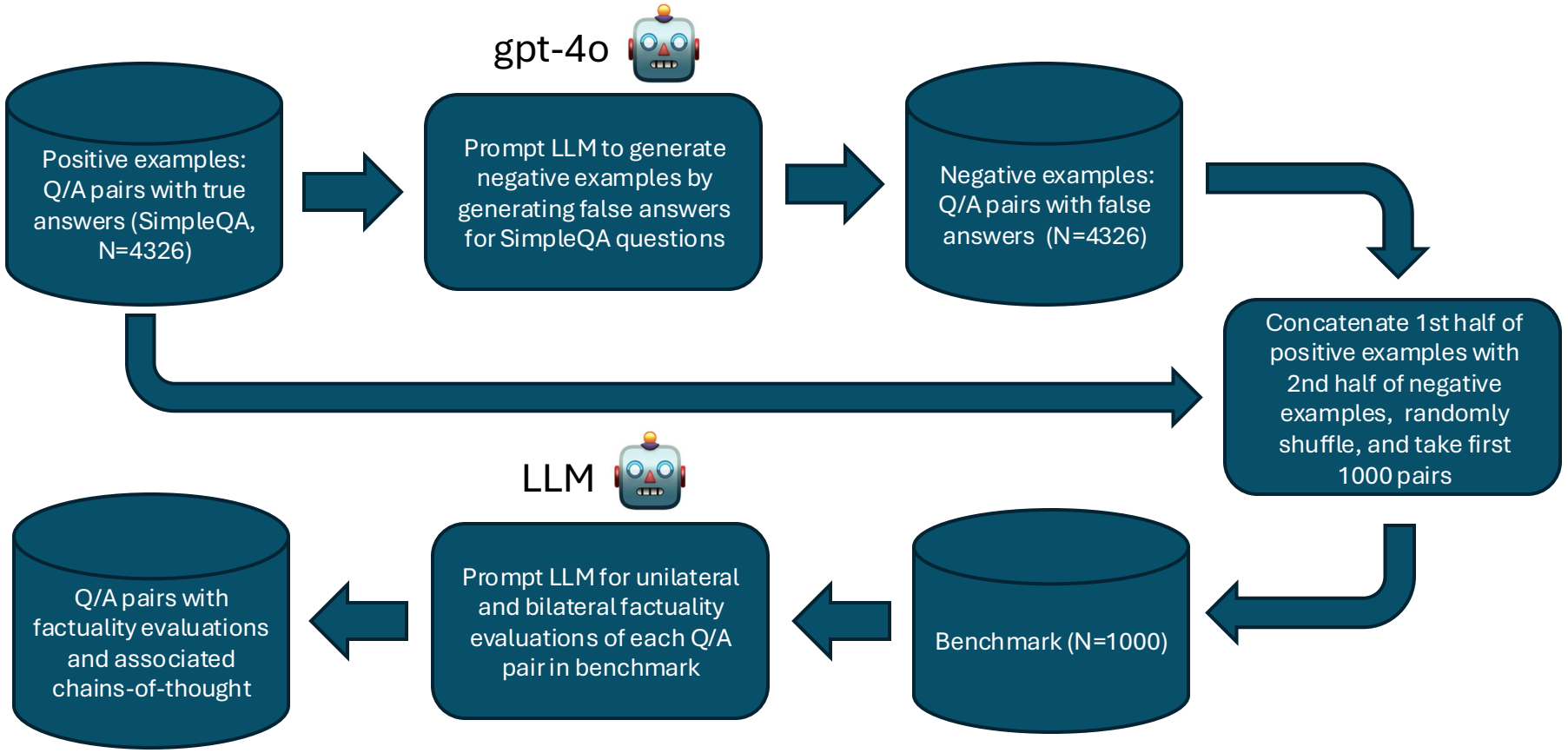
- Factuality evaluation is the assessment of whether a language model can accurately determine if a given statement or answer is factually correct (Wang et al. 2023)
  - The ability of the model to distinguish true statements from false ones
  - The model's capacity to abstain from judgment when it lacks sufficient confidence
- Benchmarks: TriviaQA (Joshi et al. 2017), Natural Questions (Kwiatkowski et al. 2019), FActScore (Min et al. 2023), FELM (Zhao et al. 2023), SimpleQA (Wei et al. 2024)
  - Most factuality benchmarks are focused on long- or short-form question answering tasks
  - Our focus is on evaluating an LLM's ability to generate a truth value for a question/answer pair

A factuality evaluation benchmark based on SimpleQA

- SimpleQA released by OpenAI in October 2024
  - 4,326 question/answer pairs
  - Questions specifically designed to be challenging for frontier models (e.g., gpt-4 and Claude Sonnet scoring <50%) in a short-form question answering (QA) task
  - Questions crafted to have single, indisputable answers and verified through multiple AI trainers
- Used to measure whether models "know what they know"
  - Also to study model calibration, measuring both explicit confidence statements and implicit confidence through repeated sampling
- We used SimpleQA to create a new benchmark for our experiments
  - Used existing SimpleQA Q/A pairs for positive examples
  - Added questions with existing SimpleQA questions paired with synthetically generated false answers for negative examples


```
{
 'metadata': {
  'topic': 'Geography',
  'answer_type': 'Place',
  'urls': [
   'https://en.wikipedia.org/wiki/Radcliffe_College',
   'https://en.wikipedia.org/wiki/Radcliffe_College',
   'https://www.braingainmag.com/7-historic-liberal-arts-colleges-in-the-us.htm',
   'https://thepeoplesarchive.dclibrary.org/repositories/2/resources/2228'
  ]
 },
 'problem': "What's the name of the women's liberal arts college in Cambridge,
Massachusetts?",
 'answer': "Radcliffe College"
}
```

Experimental framework



Bilateralism

- Bilateralism in logic (cf. Rumfitt 2000) holds that understanding a proposition requires grasping both the conditions under which it can be asserted and the conditions under which it should be denied
  - Meaning isn't just about knowing when something is true, but also explicitly understanding when it is false
- This philosophical view contrasts with unilateral approaches where only conditions for truth or assertion are primary, and falsity or denial is treated as derivative (just the absence or negation of truth)
  - Bilateralists argue this misses something fundamental about meaning and inference
- Bilateralists argue that having explicit roles for both verification and refutation leads to better logical reasoning and clearer understanding

Zero-shot CoT prompting to obtain truth values

question: answer

unilateral

Evaluate if the answer is definitively TRUE or FALSE for the given question.
Conclude EXACTLY with either:
"TRUE" - Only if every claim is independently verified
"FALSE" - If any essential claim cannot be verified
<CoT instructions>
Explain your reasoning first, then provide your conclusion.

IF output is	THEN truth value is
TRUE	t
FALSE	f
None	n

bilateral

Evaluate if the answer is definitively TRUE for the given question.
Conclude EXACTLY with either:
"TRUE" - Only if every claim is independently verified
"CANNOT DETERMINE TRUE" - If any essential claim cannot be verified
<CoT instructions>
Explain your reasoning first, then provide your conclusion.

+

Evaluate if the answer is definitively FALSE for the given question.
Conclude EXACTLY with either:
"FALSE" - Only if a contradiction is found
"CANNOT DETERMINE FALSE" - If no definitive contradiction exists
<CoT instructions>
Explain your reasoning first, then provide your conclusion.

IF outputs are	THEN truth value is
TRUE, FALSE	b
TRUE, CANNOT DETERMINE FALSE	t
CANNOT DETERMINE TRUE, FALSE	f
CANNOT DETERMINE TRUE, CANNOT DETERMINE FALSE	n

Example of bilateral evaluation (using gpt-4o)

problem	**What's the name of the women's liberal arts college in Cambridge, Massachusetts?**	
answer	**Radcliffe College**	
chains of thought	**verification**	**refutation**

	Verification Process:
	• **Parse key claims in both question and answer:**
	  ○ **Question:** The question asks for the name of a women's liberal arts college in Cambridge, Massachusetts.
	  ○ **Answer:** The answer claims that the name of the women's liberal arts college in Cambridge, Massachusetts, is "Radcliffe College."
	• **Verify each claim against known facts:**
	  ○ **Claim 1:** Radcliffe College was a women's liberal arts college.
	  ▪ **Fact Check:** Radcliffe College was indeed a women's liberal arts college, historically associated with Harvard University.
	  ○ **Claim 2:** Radcliffe College was located in Cambridge, Massachusetts.
	  ▪ **Fact Check:** Radcliffe College was located in Cambridge, Massachusetts.
	  ○ **Claim 3:** Radcliffe College is currently a women's liberal arts college.
	  ▪ **Fact Check:** Radcliffe College no longer exists as an independent women's liberal arts college. In 1999, it formally merged with Harvard University and became the Radcliffe Institute for Advanced Study, which is not a women's liberal arts college.
	• **Identify any unstated assumptions:**
	  ○ The answer assumes that Radcliffe College still exists as a women's liberal arts college, which is not true. The question does not specify whether the college is currently operational or historical, but the phrasing implies a present-day institution.
	• **Check for temporal/contextual dependencies:**
	  ○ The status of Radcliffe College has changed over time. While it was historically a women's liberal arts college, it ceased to exist in that form after 1999. The answer does not account for this temporal dependency.
	• **Validate logical connections between claims:**
	  ○ The answer logically connects Radcliffe College to the description of a women's liberal arts college in Cambridge, Massachusetts, but it fails to address the current status of the institution.
	Conclusion:
	• The answer cannot be definitively verified as true because Radcliffe College no longer exists as a women's liberal arts college, and the question does not clarify whether it is asking about a historical or current institution. Therefore, the conclusion is:
	"CANNOT DETERMINE TRUE"
		Falsification Process:
		• **Parse key claims in both question and answer:**
		  ○ **Question:** The question asks for the name of a women's liberal arts college located in Cambridge, Massachusetts.
		  ○ **Answer:** The answer claims that the name of the women's liberal arts college in Cambridge, Massachusetts, is "Radcliffe College."
		• **Search for any direct contradictions:**
		  ○ Radcliffe College was historically a women's liberal arts college in Cambridge, Massachusetts. However, it merged with Harvard University in 1999 and ceased to exist as an independent women's college. Today, it is part of the Radcliffe Institute for Advanced Study at Harvard University, which is not a women's liberal arts college.
		  ○ The question does not specify whether the college must currently exist as a women's liberal arts college. If interpreted as asking about a historical institution, the answer could be correct.
		• **Test for logical inconsistencies:**
		  ○ There is no logical inconsistency in the answer if the question is interpreted as referring to a historical women's liberal arts college. However, if the question is interpreted as asking for a currently existing women's liberal arts college, the answer would be incorrect because Radcliffe College no longer exists in that form.
		• **Check for impossible conditions:**
		  ○ It is not impossible for Radcliffe College to have been a women's liberal arts college in the past. However, it is impossible for it to currently be one, as it no longer exists as an independent institution.
		• **Identify mutually exclusive scenarios:**
		  ○ If the question is asking for a currently existing women's liberal arts college, Radcliffe College cannot be the answer because it no longer exists in that capacity.
		  ○ If the question is asking for a historical women's liberal arts college, Radcliffe College could be a valid answer.
		Conclusion:
		• The question does not explicitly specify whether it is asking for a currently existing or historical women's liberal arts college. Therefore, there is no definitive contradiction in the answer as written.
		CANNOT DETERMINE FALSE

evaluation	**n**	

Results: performance

Performance

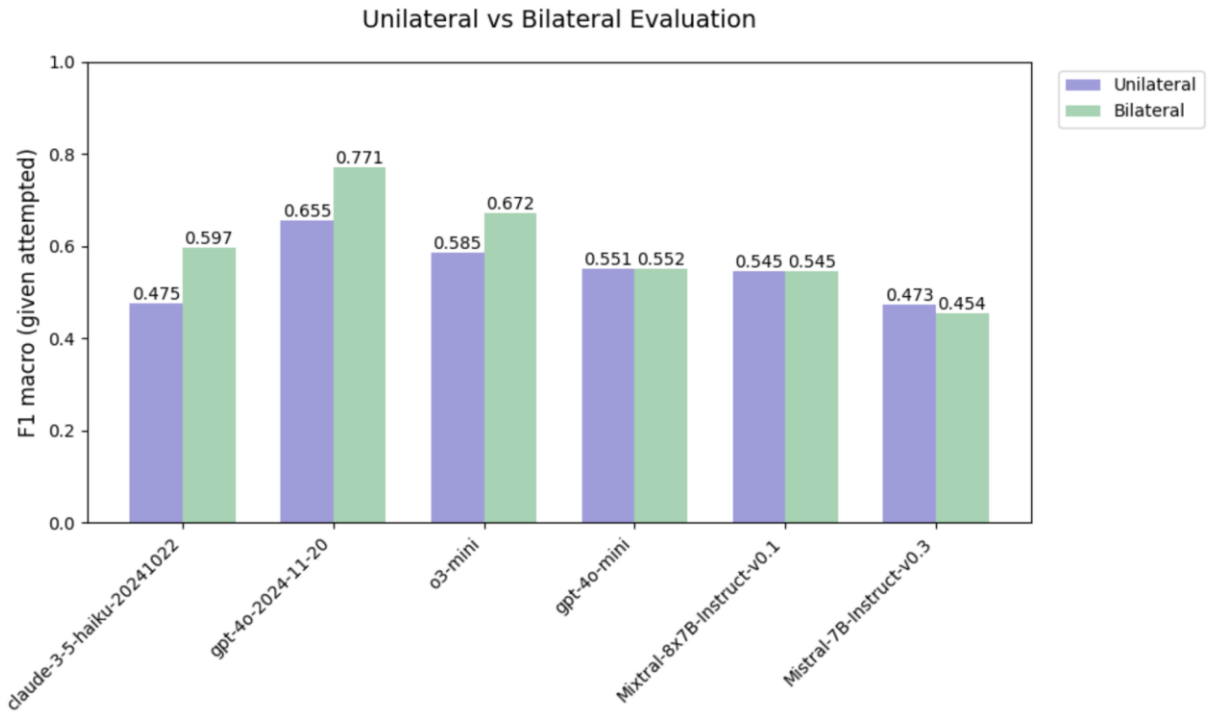
				unilateral				bilateral	
model	coverage	ACC	AUC	F1 macro	coverage	ACC	AUC	F1 macro	delta F1
claude-3-5-haiku-20241022	1.000	0.543	0.538	0.475	0.239	0.598	0.601	0.597	0.122
gpt-4o-2024-11-20	0.999	0.661	0.659	0.655	0.479	0.775	0.771	0.771	0.117
o3-mini	1.000	0.597	0.595	0.585	0.514	0.673	0.673	0.672	0.087
gpt-4o-mini	0.999	0.553	0.552	0.551	0.513	0.598	0.582	0.552	0.002
Mixtral-8x7B-Instruct-v0.1	0.992	0.545	0.546	0.545	0.474	0.557	0.556	0.545	0.000
Mistral-7B-Instruct-v0.3	1.000	0.523	0.527	0.473	0.678	0.509	0.521	0.454	-0.019

Unilateral truth value distribution

evaluation	t	n	f
model_name			
claude-3-5-haiku-20241022	0.148	0.000	0.852
gpt-4o-2024-11-20	0.374	0.001	0.625
gpt-4o-mini	0.440	0.001	0.559
mistralai/Mistral-7B-Instruct-v0.3	0.814	0.000	0.186
mistralai/Mixtral-8x7B-Instruct-v0.1	0.524	0.008	0.468
o3-mini	0.336	0.000	0.664

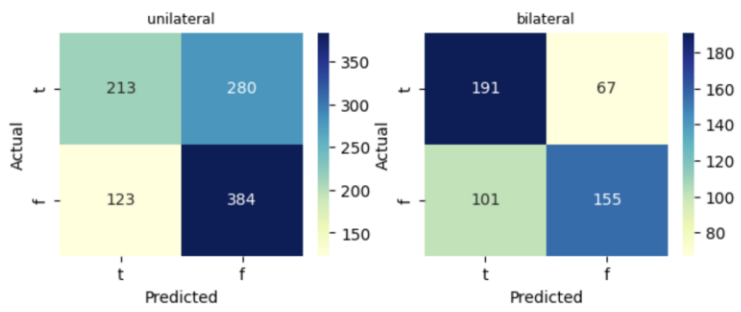
Bilateral truth value distribution

evaluation	t	n	b	f
model_name				
claude-3-5-haiku-20241022	0.122	0.730	0.031	0.117
gpt-4o-2024-11-20	0.191	0.427	0.094	0.288
gpt-4o-mini	0.407	0.398	0.089	0.106
mistralai/Mistral-7B-Instruct-v0.3	0.566	0.187	0.135	0.112
mistralai/Mixtral-8x7B-Instruct-v0.1	0.311	0.388	0.138	0.163
o3-mini	0.292	0.439	0.047	0.222

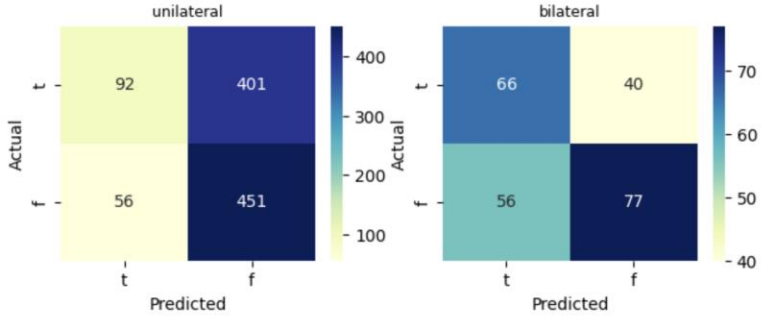


Results: confusion matrices given attempted (i.e., evaluation either t or f)

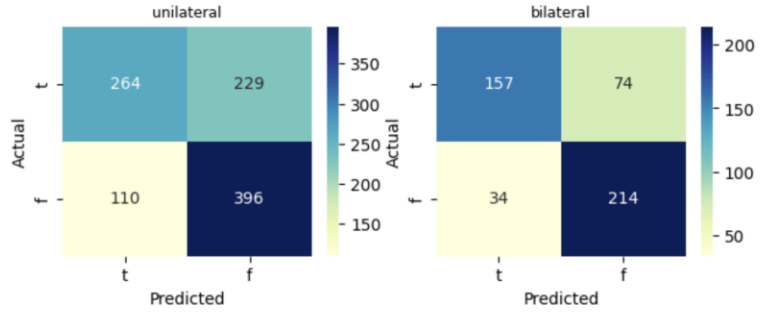
o3-mini



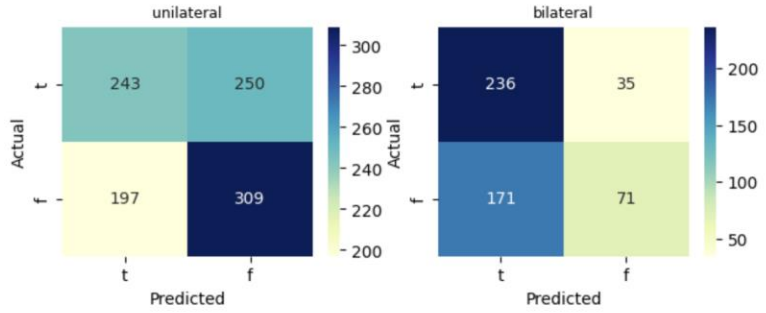
claude-3-5-haiku-20241022



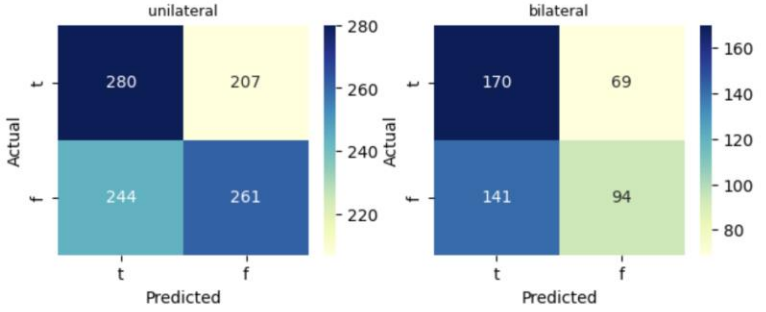
gpt-4o-2024-11-20



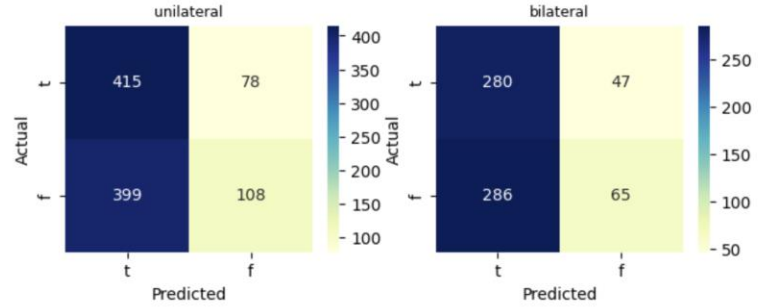
gpt-4o-mini



Mixtral-8x7B-Instruct-v0.1



Mistral-7B-Instruct-v0.3



Findings

- Bilateral evaluation generally improves factuality evaluation, particularly in frontier LLMs (8-12% improvement in F1 macro for gpt-4o, claude-3-5-haiku, o3-mini)
- The improvement comes with a coverage trade-off (24-51% coverage across all 6 LLMs evaluated)

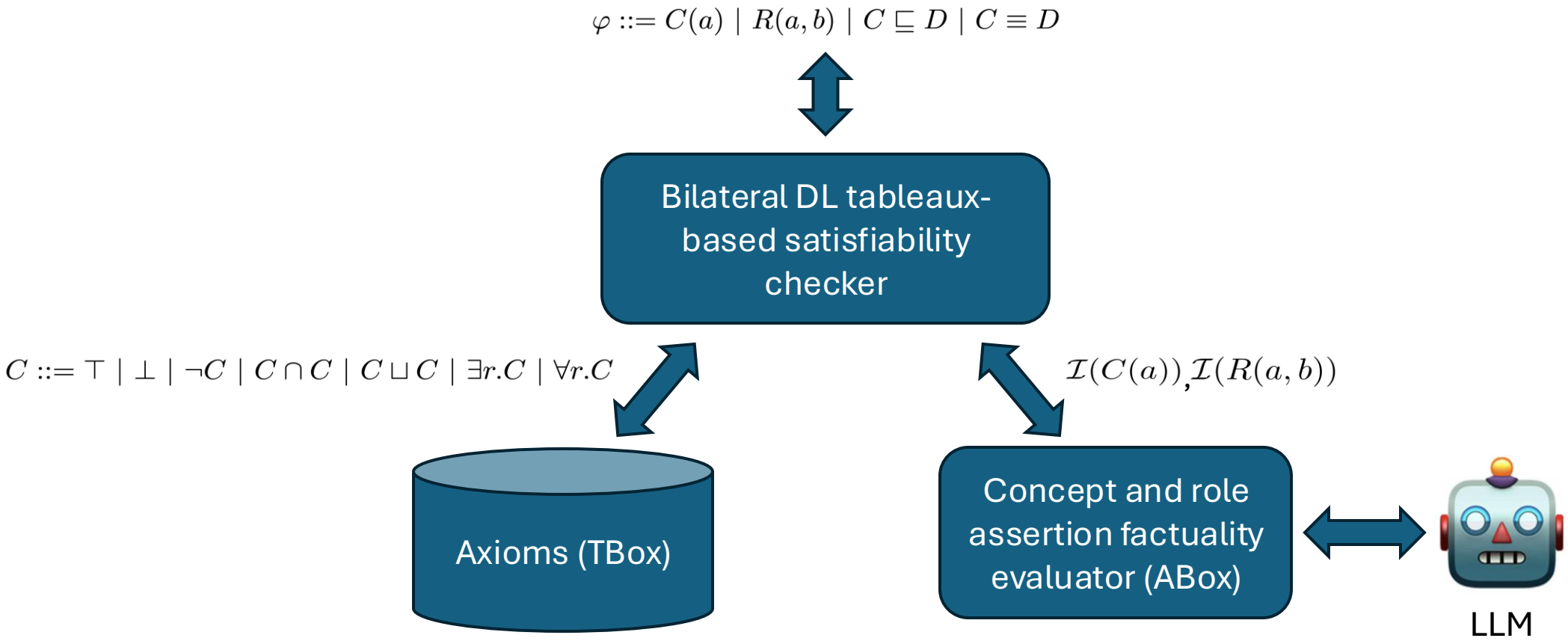
Discussion

- The coverage trade-off may be appropriate for real-world applications where abstention is preferable to incorrect evaluation
- The fact that older/smaller LLMs struggle significantly suggests they are less capable than frontier LLMs in performing the reasoning needed for bilateral factuality evaluation
- The fact that considering both verification and refutation of assertions improves factuality evaluation in frontier LLMs could be seen as providing empirical support for bilateralism in logical semantics
- This work is also potentially relevant to the questions of LLM beliefs (cf. Mandelkern & Linzen 2023, Lederman & Mahowald 2024, Herrmann & Levinstein 2024) and propositional interpretability (Chalmers 2025)

Future work

- We now have a dataset of 18,000 explanations for three factuality evaluation tasks performed over 1,000 Q/A pairs
- We will use this dataset to conduct additional experiments on
  - The effect of enhancing the LLM using content linked in SimpleQA metadata
  - The effect of repeated sampling in prompting to obtain truth values
  - LLM-as-judge approaches to hallucination (Allen & Groth 2024) and metalinguistic disagreement (Allen & Groth 2025) detection in generated chains-of-thought
- We are exploring the use of this approach to factuality evaluation to provide an LLM-grounded interpretation for a bilateral description logic (Ferguson 2021)
  - Perserves soundness and completeness of the tableaux system
  - Providing formal reasoning that is robust in the face of logical inconsistency and incompleteness in the LLM's knowledge

Using bilateral factuality evaluation in the interpretation of concept and role assertions



Thank you!

Email: b.p.allen@uva.nl

GitHub repository: https://github.com/bradleypallen/bilateral-factuality-evaluation

References

Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., ... & Zhang, Y. (2023). Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... & Petrov, S. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, *7*, 453-466.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P. W., ... & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Zhao, Y., Zhang, J., Chern, I., Gao, S., Liu, P., & He, J. (2023). FELM: Benchmarking Factuality Evaluation of Large Language Models. *Advances in Neural Information Processing Systems*, *36*, 44502-44523.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., ... & Fedus, W. (2024). Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Rumfitt, I. (2000). 'Yes and No'. *Mind*, *109*(436), 781-823.

Allen, B. P., Polat, F., & Groth, P. (2024). SHROOM-INDElab at SemEval-2024 Task 6: Zero-and Few-Shot LLM-Based Classification for Hallucination Detection. *arXiv preprint arXiv:2404.03732*.

Allen, B. P., & Groth, P. T. (2025). A Benchmark for the Detection of Metalinguistic Disagreements between LLMs and Knowledge Graphs. *ISWC 2024 Special Session on Harmonising Generative AI and Semantic Web Technologies*. CEUR-WS.org, 2025.

Mandelkern, M., & Linzen, T. (2023). Do language models refer? *arXiv preprint arXiv:2308.05576*.

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs. *Transactions of the Association for Computational Linguistics*, *12*, 1087-1103.

Herrmann, D.A. & Levinstein, B.A. (2024). Standards for belief representations in LLMs. *arXiv preprint arXiv:2405.21030*.

Chalmers, D. J. (2025). Propositional interpretability in artificial intelligence. *arXiv preprint arXiv:2501.15740*.

Ferguson, T. M. (2021). Modeling Intentional States with Subsystems of ALC. In *Proceedings of the 34th International Workshop on Description Logics (DL 2021)*. CEUR-WS.org, 2021.