

CogniBench: A Legal-inspired Framework and Dataset for Assessing Cognitive Faithfulness of Large Language Models

Xiaqiang Tang¹, Jian Li^{*2}, Keyu Hu¹, Du Nan², Xiaolong Li²,
Xi Zhang³, Weigao Sun⁴, and Sihong Xie^{*1}

¹The Hong Kong University of Science and Technology (Guangzhou)

²Hunyuan AI Digital Human, Tencent

³Beijing University of Posts and Telecommunications

⁴Shanghai AI Laboratory

Abstract

Faithfulness hallucination are claims generated by a Large Language Model (LLM) not supported by contexts provided to the LLM. Lacking assessment standard, existing benchmarks only contain “factual statements” that rephrase source materials without marking “cognitive statements” that make inference from the given context, making the consistency evaluation and optimization of cognitive statements difficult. Inspired by how an evidence is assessed in the legislative domain, we design a rigorous framework to assess different levels of faithfulness of cognitive statements and create a benchmark dataset where we reveal insightful statistics. We design an annotation pipeline to create larger benchmarks for different LLMs automatically, and the resulting larger-scale CogniBench-L dataset can be used to train accurate cognitive hallucination detection model. We release our model and dataset at: <https://github.com/FUTUREEEEE/CogniBench>

1 Introduction

With the widespread deployment of Retrieval Augmented Generation (Lewis et al., 2020; Jokinen, 2024), Large Language Models (LLMs) (Chowdhery et al., 2023; Achiam et al., 2023; Touvron et al., 2023) are increasingly expected to generate responses that adhere closely to provided context. Consequently, “faithfulness hallucination” (Huang et al.; Es et al., 2024; Saad-Falcon et al., 2024) has become a critical research problem for modern LLMs applications. Prior work (Niu et al., 2023; Mishra et al., 2024; Belyi et al., 2024) attempted to address **factual** inconsistencies, which can be directly verified by comparing the model response with the provided context.

However, as LLMs evolve, LLMs increasingly generate **cognitive statements**—such as inferences, evaluations, and explanations that extend

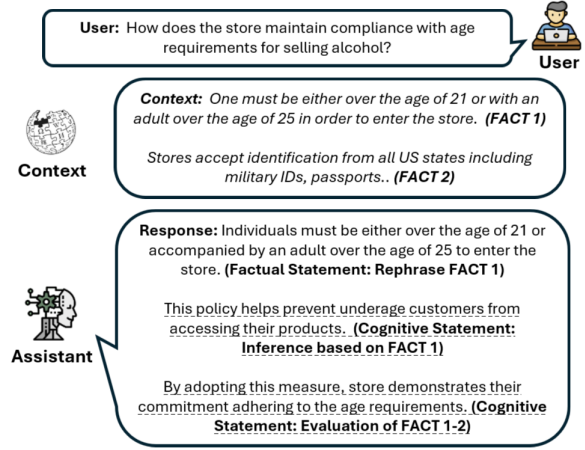


Figure 1: Difference between “factual statement” and “cognitive statement”. The former typically reproduces source materials, while the latter extends beyond the provided context to make inferences, provide explanations, or express evaluations and opinions.

beyond the verbatim context (Webb et al., 2023; Zheng et al., 2023; Xu et al., 2023). This is essential for increasing applications in medicine (Singhal et al., 2025), legislation (Singhal et al., 2025), and finance (Yao et al., 2024). For instance, a medical AI must not only recall symptoms from a patient report but also draw diagnostic conclusions. Put in a broader perspective, we categorize LLM-generated statements based on Bloom’s Taxonomy (Bloom et al., 1956): **factual statements** (recalling context) and **cognitive statements** (applying, explaining, or evaluating knowledge), as exemplified in Fig. 1. Existing benchmarks (Mishra et al., 2024; Niu et al., 2023; Hasan et al., 2021), which prioritize factual consistency, are primarily dominated by “factual statement” as shown in Fig. 2.

Assessing cognitive statements generated by LLM becomes both important and challenging: (C1) The assessment of cognitive statements lacks data and standardization. Their assessment is inherently subjective (e.g., determining whether an explanation is justified) and context-

*Corresponding Authors

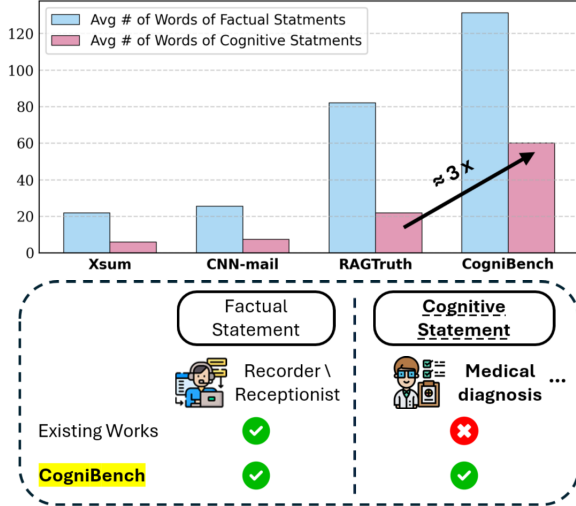


Figure 2: Dialogues in CogniBench on average contains three times more cognitive statements than previous datasets (Niu et al., 2023; See et al., 2017; Hasan et al., 2021). While factual statements mirror human roles like recorders, cognitive statements require higher-level cognitive inference beyond simple rephrasing in applications such as medical diagnosis.

dependent—requirements can vary significantly across domains (e.g., creative storytelling (Zhou et al., 2023) vs. clinical diagnosis (Singhal et al., 2025)). (C2) Manual annotation is impractical for rapidly updated LLMs, necessitating low-resource, automated assessment methods applicable to both factual and cognitive statements. We address these challenges through three key contributions:

- In response to (C1), we introduce **CogniBench** (Section 2), the first knowledge-grounded dialogue dataset and framework for assessing cognitive faithfulness. Built through rigorous manual annotation, CogniBench provides sentence-level annotations with a **legal-inspired** assessment protocol to reduce subjectivity. To accommodate diverse application requirements, we define three increasingly rigorous faithfulness criteria: *rational* (plausible but unverifiable speculation), *grounded* (contextually supported evaluations), and *conclusive* (undeniable conclusions). The increasing levels of rigor cater to the diverse needs of various LLMs’ applications, from creative storytelling (Zhou et al., 2023) to high-stakes medical diagnosis (Singhal et al., 2025).
- In response to (C2), we develop an automated annotation pipeline (Section 3) that generates *CogniBench-L*—a large-scale extension containing over 24k dialogues with sentence-level hallucination annotations. This resource supports

the training of specialized hallucination detectors and enables the systematic benchmarking of evolving LLMs.

- Empirical analysis reveals three pivotal insights: (1) cognitive statements increase with dialogue length (growing from 15% in initial turns to 50% by final turns Section 4.3), (2) LLMs exhibit **4.6× higher hallucination rates** when generating cognitive statements compared to factual ones (Section 4.4), and (3) Model reliability varies drastically—while GPT-4 produces 33.7% cognitive statements with a 60.1% hallucination rate, Gemini-Pro generates 49.9% cognitive content but suffers a 79.9% error rate (Section 4.7).

Furthermore, existing hallucination detectors struggle to assess cognitive statements accurately, with F1 dropping by 31% compared to factual cases (Section 4.6). These findings emphasize the urgency of going beyond fact-centric assessments.

2 The CogniBench Benchmark

Knowledge-grounded conversations have become a crucial scenario in which people interact with LLMs. (Lewis et al., 2020; Nakano et al.; Singhal et al., 2025). Existing datasets primarily focus on assessing the faithfulness of “factual statements”, typically assessing how well models rephrase information drawn from a given context. Moreover, the requirements for cognitive faithfulness—such as the ability to accurately make inferences, evaluations, or explanations—differ significantly across various application domains due to the diverse uses of LLMs. CogniBench aims to address these limitations by:

- Annotate a multi-round conversation dataset on the sentence level using the framework with detailed provenance information. The dataset can be used for understanding cognitive statements and detecting inconsistency (i.e. cognitive hallucinations) in LLMs.
- Establishing a legal-inspired, increasingly rigorous assessment framework, which proposes three increasingly rigorous assessment criteria. This framework minimizes annotation efforts and reduces inconsistencies, offering users the flexibility to balance creativity and faithfulness according to their specific use cases.

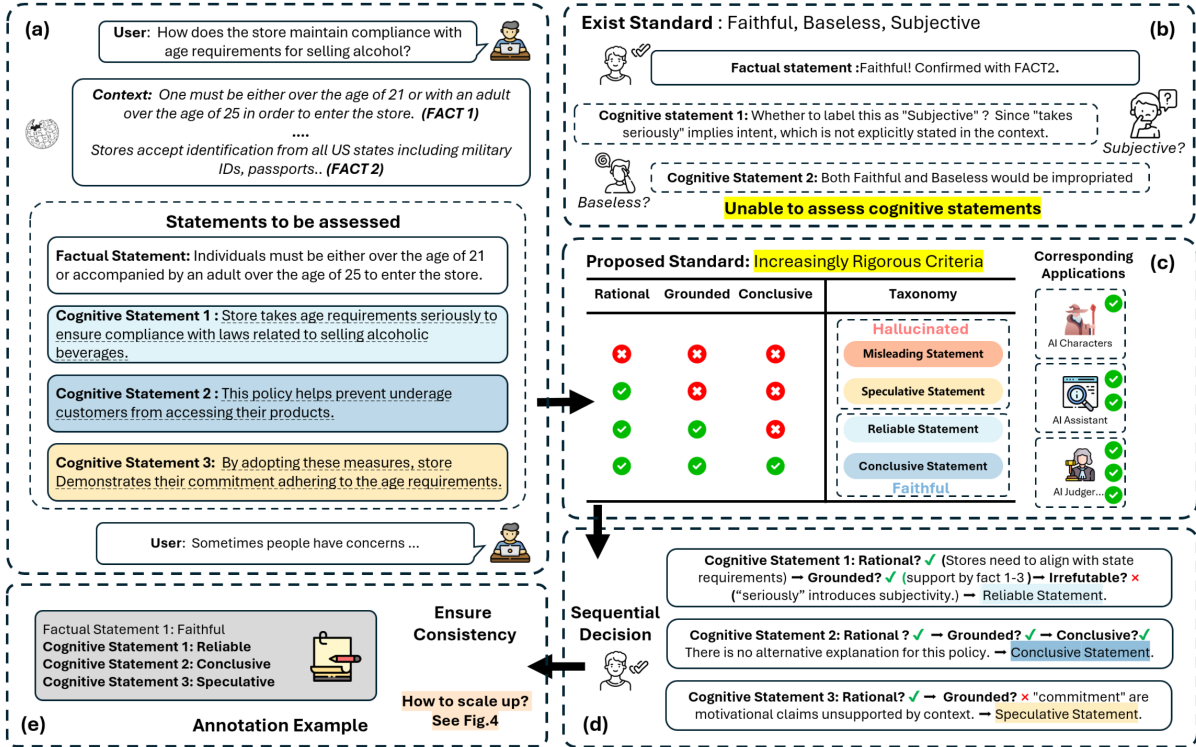


Figure 3: (a)(b): Existing faithfulness assessment standards such as “Baseless” (Niu et al., 2023) and “Subjective” (Mishra et al., 2024) are ambiguous and insufficient for assessing cognitive statements. (c): We propose three increasingly rigorous assessment criteria (i.e. **Rational, Grounded, Conclusive**) to annotate cognitive statements. (d)(e): Human annotators assess each statement in (a) by making sequential decisions. Cognitive statements can be categorized as Misleading, Speculative, Reliable, or Conclusive based on what criteria are met. (See Section 2.2 for details)

2.1 Overview of Dataset Creation

CogniBench leverages the approach from (Yang et al., 2023) to generate customized, multi-turn knowledge-grounded dialogues, minimizing factual errors. This enables us to focus on cognitive inconsistency. We use GPT-4 (Achiam et al., 2023) to generate a large corpus of dialogues, ensuring the conversations closely align with real-world LLM usage. The guidelines for annotators to assess cognitive statements are summarized as follows:

1. Identifying Statements Types: First identify whether a statement rephrases facts from context (factual statement) or the statement extends beyond the provided context, such as inference, explanation, and opinions (cognitive statement).

2. Assessing Faithfulness: For factual statements, annotators assess whether the statement is “faithful” (consistent with the provided context) or “invented” (factual hallucination not supported by provided context) (Mishra et al., 2024; Niu et al., 2023). For cognitive statements, we propose a new assessment framework where no established standard exists, as detailed in Section 2.2. Full instructions can be found in Appendix A.5.3.

2.2 A Legal-Inspired Tiered Framework for Assessing Cognitive Statements

We compare existing standards with our proposed framework in Fig. 3. This framework is characterized by two key features: 1) objectivity, ensured by pre-established laws, and 2) an increasingly rigorous design with progressively higher standards of faithfulness.

The annotations are done at the sentence level. Unlike span-level annotations of hallucinated entities or relationships of factual statements (Niu et al., 2023; Mishra et al., 2024). Cognitive statements and their cognitive nature should be assessed based on the semantics of an entire sentence.

Factual statements can be easily verified by comparing them with the provided context. For instance, Factual Statement 1 (“Individuals must be either over ...”) can be validated against FACT 1, as shown in Fig. 3 (a).

In contrast, cognitive statements extend beyond the provided context and require reasoning. Their evaluation tends to be more subjective due to their inherent nature. For example, Cognitive Statement 2 (“This policy helps prevent...”) extends FACT

2 from the provided context. Existing standards such as “baseless” or “subjective” fail to label such statements appropriately.

To build standardized assessment criteria that ensure consistent annotation, we draw inspiration from the legal domain. In legal contexts, evidence is categorized into *Direct evidence* and *Circumstantial evidence* (i.e. Indirect evidence). As illustrated in Fig. 9, *Direct evidence* refers to when a witness directly describes the original event, whereas *Circumstantial evidence* involves a witness suggesting a fact through indirect inference rather than direct observation. Similarly, cognitive statements in reasoning processes often infer conclusions from the provided context. Both require thorough analysis within their respective context to avoid misinterpretation. **We argue that the legal framework for validating *Circumstantial evidence* offers a natural analogy for assessing the faithfulness of cognitive statements.** We further elaborate on these comparisons in Appendix A.3.

Inspired by how the law validates *Circumstantial evidence*, we propose three criteria (Fig. 3 (c)) to assess the faithfulness of cognitive statements.

Criterion 1 Rational: Whether a statement is reasonably believed, even if it does not have direct evidence from the provided context. This criterion is derived from the legal distinction between *inference* with *speculation*, as outlined in Appendix A.4. Specifically, “*An inference that does not properly flow from the established fact is mere speculation.*” Virtual AI characters (Zhou et al., 2023) should adhere to this standard. Even though their primary goal may be entertainment rather than accuracy, their responses should not mislead users.

Criterion 2 Grounded: Whether a statement can be supported by the provided context. If a statement satisfies criterion 1, the annotator will proceed to assess whether the statement can be logically derived from the available information. This is established from “*An inference is a deduction of fact that may logically and reasonably be drawn from another fact or group of facts found or otherwise established in the proceedings.*”. AI assistants like WebGPT (Nakano et al.) should meet at least this standard to ensure their responses are both helpful and grounded by context.

Criterion 3. Conclusive: Whether a statement has no alternative interpretations could reasonably exist, free from subjective bias. This criterion is based on the legal requirement that “*circumstantial evidence may be introduced, but the trier-of-fact*

must be satisfied beyond a reasonable doubt that the guilt of the accused is the only reasonable inference from the proven facts” (v Villaroman, 2016). Responses from AI systems such as AI Judge, AI Doctor, or AI Trader (Yao et al., 2024; Singhal et al., 2025; Ding et al., 2024) should meet this standard. These systems make high-stakes decisions that directly impact human lives and livelihoods. Therefore, LLMs must make conclusive statements to minimize the risk of error or bias.

With these three criteria, a statement can be categorized into one of the following: **Misleading, Speculative, Reliable, or Conclusive.** For example, Cognitive Statement 1, labeled as Reliable (which might have been categorized as subjective or a hallucination under previous standards) is both Rational and Grounded, yet contains subjectivity. Despite this, it remains useful for AI assistant applications. Additional annotated examples are presented in Fig. 10.

Annotation Method	IAA	QA Instances
Independent Multi-Class Classification	91.51%	25 (15 real-time QA + 10 post-hoc feedback)
Sequential Decision Framework	96.19%	13 (6 real-time QA + 7 post-hoc feedback)

Table 1: Inter-Annotator Agreement (IAA) and Quality Assurance (QA) Effort (per 500 statements): Our Sequential Decision Framework achieves higher degree of agreement with less annotation effort.

Our framework organizes the annotation process into a sequential decision-making structure (see full annotation protocol in Appendix A.5.2), guiding annotators through a sequential decision-making process to classify statements. The design ensures the annotator only considers a more rigorous criterion (e.g., grounded or conclusive) once previous less rigorous conditions (e.g., rational) are met, reducing cognitive load and the potential for conflicting interpretations. As shown in Table 1, this approach outperforms independent classification (i.e., direct ask annotator to categorize a statement into one type such as misleading, reliable) , achieving a 96.19% inter-annotator agreement (IAA) while also reducing QA instances by 48%.

3 Automated Data Expansion

The rapid emergence of new LLMs poses significant challenges for faithfulness assessment, as manual annotation is time-consuming and costly. Moreover, relying on small-scale datasets is insufficient for the fine-tuning a generalized hallucination detection model (Niu et al., 2023). While synthetic

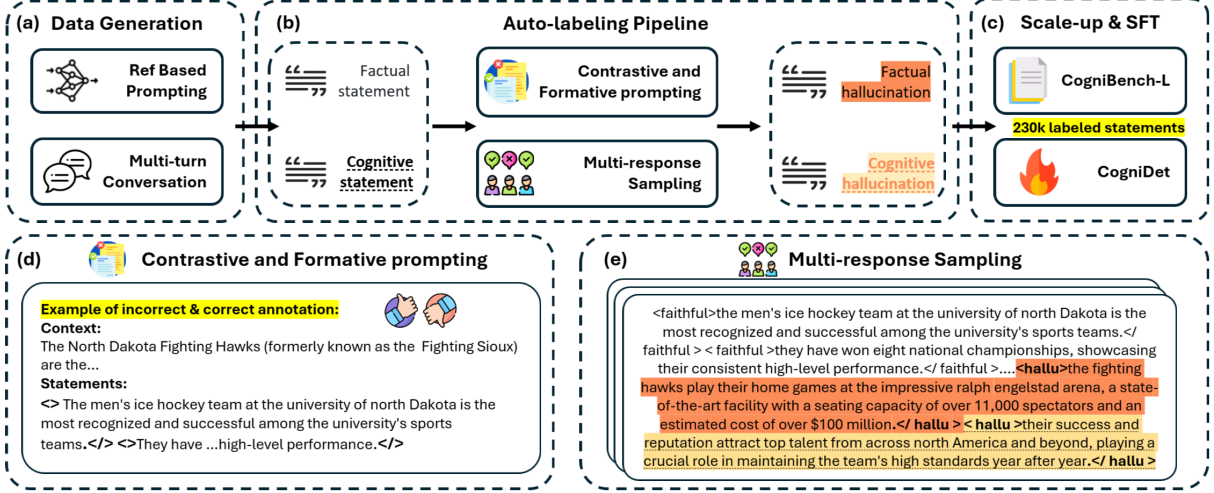


Figure 4: Auto-Labeling pipeline for hallucination detection. (a): We generate multi-turn, knowledge-grounded conversations using (Yang et al., 2023). (b): A trained model classifies statements as factual or cognitive. Two prompting strategies are employed to generate sentence-level hallucination annotations with GPT-4. (c): This process results in a large-scale hallucination dataset CogniBench-L. We fine-tune an detection model to identify hallucinations in both factual and cognitive statements.

hallucination texts (Mishra et al., 2024; Luo et al., 2024) can guide LLMs to generate a considerable amount of annotated hallucinated data, they create a gap between simulated and real-world scenario.

To address this, we propose a fully automated labeling method that utilizes LLMs as judges to label dialogue data generated by advanced LLMs. This cost-effective approach can assess the faithfulness of new models and provides an scalable way to expand the CogniBench dataset into CogniBench-L, a comprehensive training corpus for hallucination detection model.

Results presented in Table 5 show that Auto-Labeling can matches human annotators’ accuracy making it a reliable proxy for assessing new LLMs. Additionally, Using the large-scale CogniBench-L dataset, we fine-tuned an 8B model CogniDet, which achieves state-of-the-art performance in detecting hallucinations in both factual and cognitive statements, as shown in Section 4.6. CogniDet is effective for low-cost hallucination detection in daily applications.

3.1 Contrastive and Formative Prompting

Achieving reliable sentence-level annotation of LLM outputs presents two primary challenges: (1) enabling the annotation LLM to accurately understand the standards outlined in Section 2.2, and (2) the high cost of annotating sentences individually within dialogues.

Building on in-context learning (Brown, 2020; Dong et al., 2022), we provide the LLM with annotated examples that demonstrate both the desired

task and the expected output format. Unlike conventional labeling tasks, internalizing our faithfulness criteria is complicated by the inherent ambiguity of natural language. To address this, we employ a two-step process:

- 1. Initial Prompting and Diagnosis:** We first prompt LLM with annotation examples from human annotators and ask the LLM to annotate statements accordingly. We then identify and collect common annotation errors. For example, when labeling factual statements, the LLM may misclassify world-knowledge-verifiable facts not mentioned in the context as faithful. When labeling cognitive statements, it may mistakenly classify logically deduced conclusions (e.g., Cognitive Statement 2 in Fig. 3 (a) “This policy helps prevent underage customers from accessing their products”) as hallucinations, even if they are contextually justifiable.

- 2. Contrastive Examples:** Based on these observations, we provide positive and negative examples in the prompt to clarify the boundaries between faithful statements and hallucinated statements.

To avoid labeling sentences individually, we implement a formative prompting approach that guides the LLM to annotate every sentence in a single response. We use the NLTK toolkit (Bird et al., 2009) to tokenize the text into sentences, then enclose each sentence with HTML-like markers ‘<>’ and ‘</>’. The LLM is then instructed to place the assessment for each sentence within these markers (e.g., <faithful>, <hallu>). Prompts are shown in Appendix A.7. For cognitive statements, a second-stage prompt further categorizes

them into four sub-categories.

3.2 Multi-response Sampling

Although Contrastive and Formative Prompting can significantly reduce the ambiguity in defining cognitive faithfulness, individual LLM responses may still manifest hallucinatory judgments. To further enhance reliability, we adopt a sampling-based technique inspired by recent hallucination detection methods (Farquhar et al., 2024).

In practice, we prompt the LLM five times for each annotated instance, maintaining identical prompts. This yields multiple independent assessments for the same sentence. We then apply a majority-vote criterion: a sentence is deemed faithful only if the majority of sampled responses classify it as such. By aggregating multiple judgments, this approach mitigates the impact of occasional hallucinations or outlier predictions, ensuring that the final annotated dataset more accurately reflects credible cognitive statements.

As shown in Table 5, our contrastive and formative prompting (CFP) and Multi-response Sampling (Sampling) significantly enhance the LLM’s ability to discern faithful statements from hallucinations.

4 Experiments

4.1 Matrices and Baseline

We report precision, recall, and the F1 score for detection results. For instance: $Recall = \frac{\# \text{ of Words in detected hallucination sentence}}{\# \text{ of Words in labeled hallucinated sentence}}$. Precision and F1 scores are calculated using analogous formulas. Using CogniBench, we conducted experiments with the following six distinct algorithms for hallucination detection as detailed in Appendix A.8.1.

4.2 Statistic results of Cognibench

We present basic statistics of the CogniBench dataset in Table 2. CogniBench is a human-annotated following the framework proposed in Section 2.1. It is the first benchmark specifically designed to assess the faithfulness of cognitive statements while also incorporating factual statement annotations. Unlike previous datasets (Niu et al., 2023; Mishra et al., 2024), CogniBench uniquely features multi-turn conversations with extensive and diverse contexts, enabling a comprehensive assessment of language models’ cognitive reasoning.

Due to the scalability limitations of human annotation, we employ GPT-4 to generate CogniBench-

L, an automatically labeled dataset, using the methodology outlined in Section 3. CogniBench-L is 100 times larger than CongiBench. This expansion facilitates the development of a more generalized and robust hallucination detection model. As demonstrated in Table 3, models trained on CogniBench-L outperform existing faithfulness-based hallucination detection approaches, underscoring the effectiveness of our dataset in enabling improved assessment and mitigation of cognitive hallucinations in large language models.

4.3 Cognitive Dynamics in Conversation

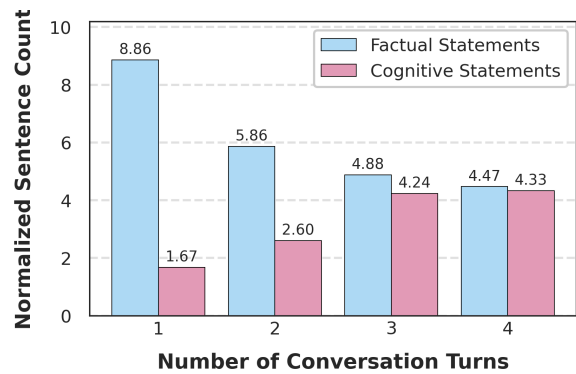


Figure 5: Average number of factual statements decreases as the number of conversation turns increases, while the number of cognitive statements increases with more turns.

Cognitive statements increase with the length of conversation In Fig. 5, we illustrate the dynamics of factual and cognitive statements relative to the number of conversation turns.

The distribution of factual and cognitive statements across conversation turns reveals distinct patterns. LLMs initially follow the provided context closely and generate factual statements, but as the dialogue progresses, they produce more cognitive statements. Showing a shift toward deeper, more reflective interactions as the conversation unfolds.

This pattern underscores the significance of examining multi-turn, knowledge-grounded conversations, where the role of cognitive statements becomes more pronounced as the exchange develops. Current datasets, however, are either limited to single-turn conversations (Niu et al., 2023) or are not long enough to capture cognitive statements effectively (Luo et al., 2024).

4.4 Statement Distribution

LLMs exhibit a higher risk of hallucination when generating cognitive statements compared to factual ones. We analyze the distribution of

Dataset	Num Response	Num Conversation	Num Labeled Sentences	Num Context Words (min-max (avg))	Words per Response
CogniBench	264	179	2516	297–1252 (696.94)	50-432(200.44)
CogniBench-L	24084	7058	234164	8-1409 (711.71)	8-709(201.38)

Table 2: The basic statistics of CogniBench.

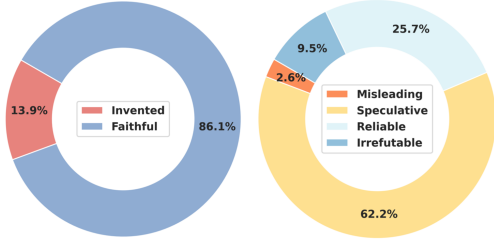


Figure 6: **Left:** Distribution of Factual statements. **Right:** Distribution of Cognitive statements. **LLMs are more likely to generate cognitive hallucinations (i.e., misleading and speculative statements) when producing cognitive statements.**

statements in CogniBench according to their faithfulness. For factual statements, 13.9% were found to have no supporting sources from the context and were thus labeled as “invented” by our annotators. In contrast, for cognitive statements, 62.2% were categorized as speculative, satisfying only Criterion 1 in Section 2.2. Furthermore, 2.6% of cognitive statements did not meet Criterion 1 and were categorized as misleading. Consequently, the overall hallucination rate for cognitive statements is 64.8%, compared to 13.9% for factual statements.

These findings highlight a key limitation of LLMs: while they are generally accurate in recalling and understanding factual content, they often struggle to apply this knowledge in cognitively demanding tasks, underscoring the urgent need to improve their faithfulness in such contexts.

4.5 Faithfulness Dynamics in Conversation

In Fig. 7, we examine the faithfulness hallucination occurrence positions. As shown in the top plot, both factual and cognitive hallucinations are significantly more likely to occur as the number of conversation turns increases. This observation is consistent with recent studies (Niu et al., 2023; Liu et al., 2024; Wang et al., 2024), which suggest that the faithfulness of LLMs tends to decrease as the context lengthens. CogniBench further demonstrates that, as conversations progress, LLMs exhibit a consistent decline in faithfulness.

Additionally, the bottom plot presents the distribution of hallucinations within individual conversation turns. We observe that factual hallucinations

tend to occur in the middle of a turn. In contrast, cognitive hallucinations are more likely to occur at the end of a turn, with some also happening at the beginning. This pattern suggests that LLMs are more prone to introducing questionable conclusions or opinions at the beginning or end of a turn where hallucinations are especially likely to occur.

4.6 Evaluate Hallucination Detection Methods on Cognibench

	Method	Overall	Factual Hallucination	Cognitive Hallucination
Prompting	ChatGPT-3.5	48.54	22.98	56.57
	ChatGPT-4	58.03	46.82	66.04
NLI	Tasksource (COLING 2024)	26.87	27.10	26.75
	SelfCheckGPT (EMNLP 2023)	45.81	32.08	61.10
E2E	Fava (CoLM 2024)	7.90	12.90	5.10
	RAGTruth (ACL 2024)	23.90	45.30	11.20
Ours	Auto-Labeling	82.20	82.50	81.90
	CogniDet 8B	70.30	64.40	73.80

Table 3: Hallucination detection performance (sentence-level F1) on CogniBench. Cognitive hallucination refers to misleading and speculative statements. Overall scores represent macro-averages across both categories.

We evaluate existing hallucination detection methods on CogniBench, revealing significant performance gaps between factual and cognitive hallucination detection. While Fava and RAGTruth achieve 12.9% and 45.3% F1 on factual hallucinations, respectively, their performance drops to 5.1% and 11.2% for cognitive hallucinations, highlighting CogniBench’s challenging nature.

Auto-Labeling pipeline (Section 3) achieve performance with 82.2% overall F1, closely matching human annotation and demonstrating reliable annotation capability. We fine-tune a 8B-parameter model CogniDet on the auto-labeled CogniBench-L corpus which reaches 70.3% F1 with a single forward pass—far cheaper than NLI baselines that perform pairwise sentence–context comparisons. Complete detection examples are provided in Appendix A.10. We believe auto-Labeling pipeline can serve as a proxy to assess new LLMs, and CogniDet is suitable for low-cost hallucination detection in everyday applications.

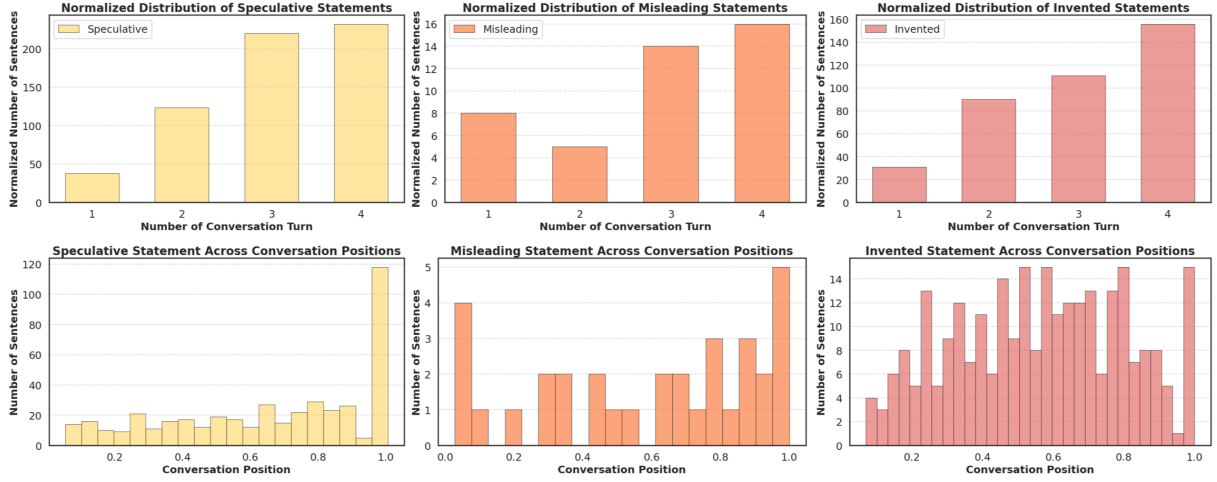


Figure 7: **Top:** Hallucinated statements increase with more conversation turns. **Bottom:** Distribution of hallucinated statements across dialogues. **Cognitive hallucinations (speculative or misleading statements) tend to occur at the beginning or end of a dialogue, while factual hallucinations are more likely to appear in the middle of the dialogue.**

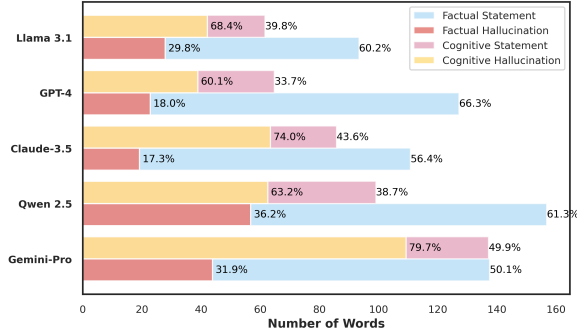


Figure 8: Evaluation of factual and cognitive statement portions with respect to their faithfulness via the proposed auto-labeling pipeline. Models evaluated: Llama 3.1 (Llama 3.1 70B), GPT-4 (GPT-4-1106-preview), Claude 3.5 (Claude-3.5-sonnet-20241022), Qwen 2.5 (Qwen2.5-72B-Instruct), and Gemini-Pro (Gemini-1.0-Pro).

4.7 Evaluate the Faithfulness of LLMs

LLMs show distinct preferences for generating cognitive versus factual statements, and the reliability of their cognitive statements varies. To evaluate this, we applied the proposed auto-labeling method to five popular advanced models, both closed-source and open-source, as shown in Fig. 8. All models were prompted with the same input and followed the approach in (Yang et al., 2023) for generating knowledge-grounded conversations.

Results show that GPT-4 generates the largest portion of factual statements, with 66.3% of its response categorized as factual. On the other hand, Gemini-Pro tends to steer the conversation toward generating more cognitive statements, accounting for 49.9% of its response. In terms of faithfulness, GPT-4 exhibits the highest reliability for cogni-

tive statements but still suffers from hallucinations, with 60.1% of cognitive statements identified as hallucinations. Claude-3.5 stands out for its high faithfulness in factual statements, with only 17.3% of those statements being hallucinations.

5 Conclusion

In this paper, we introduced a legal-inspired framework and **CogniBench**, a novel dataset for assessing the faithfulness of LLMs in generating factual and cognitive statements. We establish an increasing level of rigorousness evaluation framework, catering to application-specific requirements. We have analyzed of LLMs’ dialogue patterns and the dynamics of their faithfulness. We found that while LLMs are generally capable of rephrasing factual information accurately, their reliability drops significantly when generating cognitive statements. This highlights the need to improve cognitive capabilities in LLMs.

Additionally, we propose an automated labeling method using LLMs as judges, facilitating access to a broad range of popular models. This method allows for the scalable expansion of **CogniBench-L**. We fine-tuned **CogniDet** on CogniBench-L, which achieved state-of-the-art performance in detecting hallucination in both factual and cognitive statements. We believe this work lays the groundwork for the future development of LLMs’ cognitive capabilities, enabling more intelligent and safer deployment in applications.

Limitations

The study of hallucination in large language models (LLMs) is rapidly evolving. Our work presents a comprehensive framework for assessing the faithfulness of both factual and cognitive statements in LLMs. However, several caveats remain.

Domain specificity. Our benchmark centres on common-sense, open-domain dialogues. High-stakes applications (e.g., medical, financial, national-security) demand annotators with domain expertise and finer-grained taxonomies of error. Extending our protocol to such domains will therefore require specialised guidelines and expert-curated test sets.

Legal analogy. Legal concepts vary across jurisdictions, jurisprudence, and judicial discretion. We employ legal framework to structure cognitive-faithfulness judgements, not to impose rigid, jurisdiction-specific standards. The analogy should be read as inspiration for systematic reasoning, not as a normative legal framework.

Source coverage and bias. Wikipedia offers broad coverage and a transparent revision history, making it a convenient starting point; nevertheless, its editorial biases may skew reference distributions. Because our toolkit is fully open-source and modular, future iterations can plug in alternative corpora to mitigate such biases.

Ethical considerations

This work is in full compliance with the Ethics Policy of the ACL. We acknowledge that responses generated by LLMs in this study may contain inaccuracies. Aside from this, to the best of our knowledge, there are no additional ethical issues associated with this paper.

Acknowledgments

Sihong Xie was supported by the Department of Science and Technology of Guangdong Province (Grant No. 2023CX10X079), the National Key R&D Program of China (Grant No. 2023YFF0725001), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), and Education Bureau Guangzhou Municipality.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2024. Luna: An evaluation foundation model to catch language model hallucinations with high accuracy and low cost. *arXiv preprint arXiv:2406.00975*.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Benjamin S Bloom et al. 1956. Taxonomy of. *Educational Objectives*.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.
- Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Bryan A Garner et al. 2004. Black's law dictionary.
- Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. *The llama 3 herd of models*. Preprint, arXiv:2407.21783.

- Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.
- Kristiina Jokinen. 2024. The need for grounding in llm-based dialogue systems. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge)@ LREC-COLING-2024*, pages 45–52.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Wolters Kluwer Law et al. 2008. *Criminal law*. Wolters Kluwer.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12.
- Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *arXiv preprint arXiv:2406.07070*.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.
- Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- New York State Unified Court System. n.d. [Guide to new york evidence](#). Accessed: 2025-02-01.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.
- Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
- Damien Sileo. 2024. [tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- R v Villaroman. 2016. [R v villaroman](#). 2016 SCC 33 (CanLII), [2016] 1 SCR 1000, per Cromwell J, at para 55. Supreme Court of Canada.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3712–3724.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Wikimedia Foundation. 2025. English wikipedia dump. <https://dumps.wikimedia.org/enwiki/>. Accessed: 2025-05-27.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. Re-fgpt: Dialogue generation of gpt, by gpt, and for gpt. *arXiv preprint arXiv:2305.14994*.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pages 108–112.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Character-glm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

A Appendix

A.1 Related Work

A.2 Hallucination of Large Language Models

Hallucination in LLMs can be generally categorized into two types: factuality hallucination, where generated content deviates from established world knowledge (e.g., claiming “Mars has oceans”), and faithfulness hallucination, where the generated response is inconsistent with the provided context (e.g., misrepresenting a source document’s information). (Huang et al.)

Existing research has demonstrated that incorporating up-to-date, relevant knowledge in the prompt can effectively reduce factuality hallucination (Lewis et al., 2020). In contrast, faithfulness hallucination is persistent between the model’s response and the provided context.

Early work focused on summarization tasks (See et al., 2017; Hasan et al., 2021), but the rise of RAG systems (Lewis et al., 2020) has shifted attention to faithfulness in knowledge-grounded generation, where hallucination risks compound with contextual complexity.

A.2.1 Faithfulness Hallucination Datasets

Recently, (Mishra et al., 2024) proposed a fine-grained taxonomy for faithfulness hallucinations in long-form text generation by synthesizing hallucinated examples. (Niu et al., 2023) developed a word-level hallucination detection method specifically for RAG applications, leveraging manually annotated data. However, existing datasets prioritize “factual statements,” focusing on surface-level errors such as entity mismatches or paraphrasing inaccuracies. Moreover, they are typically single-turn dialogues (Niu et al., 2023) or short contexts dominate (Luo et al., 2024), neglecting multi-turn interactions where cognitive statements emerge (Section 4.3).

CogniBench addresses these gaps by curating dialogues rich in cognitive statements—claims requiring inference (“This policy prevents underage access”), evaluation, or hypothetical reasoning (Figure 1). These statements reflect real-world LLM applications in medicine, law, and finance, where unfaithful reasoning carries high stakes.

We argue that as LLMs continue to advance, the importance of “cognitive statement” will rise, and users will increasingly rely on their capabilities for complex reasoning and decision-making.

A.2.2 Faithfulness Hallucination Detection

Faithfulness assessment frameworks span tasks from summarization (Fabbri et al., 2022; Maynez et al., 2020; Scialom et al., 2021) to knowledge-grounded dialogue (Niu et al., 2023; Mishra et al., 2024; Manakul et al., 2023). However, these methods prioritize factual consistency (e.g., detecting entity/relation errors) and struggle with cognitive statements. For instance, an LLM might faithfully cite a medical report’s data (factual) yet draw an unfounded diagnostic conclusion (cognitive). Assessing the latter requires reasoning about contextual

plausibility, not just textual overlap—a challenge existing tools are unequipped to address.

Based on CogniBench, we developed a detection model that excels at detecting both hallucinations in "factual statements" and "cognitive statement" Table 3.

A.3 The Equality of Assessing Cognitive Statements and the validation of Circumstantial Evidence

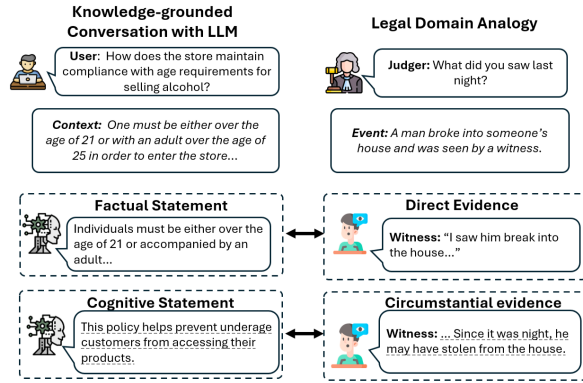


Figure 9: Drawing parallels between legal domain evidence and Cognitive Statements in knowledge-grounded conversations with the LLM

As shown in Fig. 9, the process of evaluating cognitive statements in knowledge-grounded conversations can be viewed through the lens of legal domain evidence. In this analogy, the user assumes a role akin to a judge in a legal proceeding. The context provided by the conversation serves as the event that has occurred.

Factual statements within a conversation directly align with *Direct evidence* in a legal context. For example, when the user asks about age requirements in the store, a factual statement like "Individuals must be either over the age of 21 or accompanied by an adult" can be directly validated by comparing it with the context provided. This mirrors how direct evidence in a legal case directly points to the truth of the matter.

On the other hand, cognitive statements require more nuanced assessment. Just as *Circumstantial evidence* suggests facts based on indirect observation, cognitive statements extend beyond the given context and rely on logical inferences. In the figure, cognitive statements like "This policy helps prevent underage customers from accessing their products" rely on reasoning drawn from the provided context, just as circumstantial evidence in a court case requires drawing reasonable inferences to support a

conclusion.

By leveraging the long-established and reliable standards from the legal domain, we can ensure a more rigorous and systematic assessment of cognitive statements in knowledge-grounded conversations with LLMs.

A.4 Definitions

We refer to (Garner et al., 2004; New York State Unified Court System, n.d.; Law et al., 2008) for following definition:

(1) *Direct evidence* is evidence of a fact based on a witness’s personal knowledge of that fact acquired by means of the witness’s senses. *Direct evidence* means evidence which immediately points to the question at issue.

(2) *Circumstantial evidence* is indirect evidence that does not, on its face, prove a fact in issue but gives rise to a logical inference that the fact exists. *Circumstantial evidence* requires drawing additional reasonable inferences in order to support the claim.

(3) *Drawing inferences* can be described as a two-step process. The first step is to find that the facts from which the inference is to be drawn have been proven in the trial. If not then any inference is of necessity nothing more than speculation. The second step is to make an inference from the proven facts that is reasonable, rational and logical

(4) *Inference vs. Speculation*: An inferred fact must be one that is "reasonably and logically drawn from a fact or group of facts established by the evidence." An *inference* that does not properly flow from the established fact is mere *speculation*.

An *inference* is a "deduction of fact that may logically and reasonably be drawn" from objective facts. *Speculation* is when the judge theorizes without evidentiary support or where a conclusion is drawn in the absence of a proven fact.

A.4.1 Examples of Direct and Circumstance Evidence

In a trial, if someone is trying to prove it rained on a certain morning, a witness can provide evidence in two ways.

Direct evidence: If the witness says they walked to the subway, saw rain, felt it, and heard it, this is direct evidence that it rained that morning.

Circumstantial evidence: If the witness says it was clear when they walked to the subway, but later saw people on the train with wet umbrellas and

clothes, this doesn't directly prove it rained. However, it suggests, through logical inference, that it likely rained, making it circumstantial evidence.

A.5 Details of CogniBench Annotation

A.5.1 Construction of Knowledge-grounded Dialogues

Knowledge source. All background passages are retrieved exclusively from the 2025 Wikipedia (Wikimedia Foundation, 2025).

Reference selection. Every candidate passage is first tagged with a topic label produced by LLAMA-3-70B-INSTRUCT. We then draw an equal number of passages from each label to guarantee balanced topical coverage (see Fig. 11).

Dialogue generation. We leverage the RefGPT approach (Yang et al., 2023) to generate customized, multi-turn, knowledge-grounded dialogues while minimizing factual errors. User prompts are capped at ~150 words, and assistant replies at ~500 words—limits that preserve context-window budget yet allow fully referenced answers. Dialogue turns are authored by a mixture of state-of-the-art instruction-tuned models, with GPT-4-1106-PREVIEW serving as the primary engine alongside other large language models to ensure realistic and coherent interactions.

A.5.2 Annotation Protocol

Identifying hallucinations within cognitive statements is a challenging task that requires critical thinking and a deep understanding of the logical flow of text across various topics. To ensure the reliability of the annotation process, we have implemented the following measures:

- We hired annotators from a professional vendor. All annotators are proficient in English and possess at least a bachelor's degree.
- Annotators undergo comprehensive training to ensure they fully understand the annotation standards outlined in Section 2.2. They must also pass a test annotation set before beginning formal annotation.
- Each sentence in CogniBench is reviewed by two annotators; in case of discrepancies, a third review is conducted.
- An online feedback and QA form is provided, allowing annotators to raise any questions or clar-

ifications with us if they encounter any confusion during the annotation process.

Each annotator gets paid at a rate of 0.30\$ per statement assessed.

A.5.3 Full Instruction for Annotator

As shown in Table 4, we provide comprehensive instructions for annotators to follow when labeling each dialogue. The instructions include steps for identifying irrelevant statements, classifying the type of statements, and assessing factual and cognitive statements. Additionally, we offer positive and negative examples to help annotators better understand the guidelines for each category. For simplicity, these examples are omitted here.

A.5.4 Examples of Annotation

We provide annotated examples in Fig. 10 for four types of cognitive statements: Misleading, Speculative, Reliable, and Conclusive. The context has been excerpted for ease of reading.

A.6 Topic Distribution

As shown in Fig. 11, to enhance the evaluation of large language models, we collect a diverse set of topics from Wikipedia, ensuring broad representation of real-world use cases in the knowledge-grounded multi-turn dialogues. This strategy allows us to rigorously test the generalization abilities of models evaluated on CogniBench.

A.7 Prompt used to generate CogniBench-L

We present the prompt used to generate CogniBench-L, as detailed in Table 6. The user prompt is structured to first provide examples of both incorrect and correct annotations, covering both cognitive and factual statements. These examples are drawn from the Initial Prompting and Diagnosis process, where we identified recurring annotation patterns. The conversation is segmented into single-turn dialogues, with each sentence enclosed in `< >` and `</>` to facilitate sentence-level annotation. This structure allows the LLM to focus on annotating individual sentences. The sample prompt is then passed to GPT-4 five times, and a majority-vote approach is applied to determine the final annotation for each sentence.

A.8 Implementation Detail

A.8.1 Hallucination Detection Methods Setup

Using CogniBench, we conducted experiments with the following six distinct algorithms for hallu-

STEP 1: IDENTIFY IRRELEVANT STATEMENTS
A statement is irrelevant if it contains no meaningful information related to the dialogue context or task.
STEP 2: CLASSIFY STATEMENT TYPE
Factual Statement: Makes claims about objective facts (e.g., dates, events, entities). Verifiable by directly comparing with the provided context (e.g., retrieved documents, dialogue history). Example: "stores accept various forms of unexpired identification, including ids from all us states."
Cognitive Statement: Involves reasoning, interpretation, opinions, predictions, or subjective descriptions. Requires inference from context or indirect evidence. Example: "This practice ensures that they verify the age of their customers accurately and consistently"
STEP 3: EVALUATE FACTUAL STATEMENTS
Faithful: Facts are supported by the context; no contradictions. Invented: Otherwise.
STEP 4: EVALUATE COGNITIVE STATEMENTS
Apply the following rules in sequential order: Rule 1: Rational: Whether the statement is plausible speculation. Rule 2: Grounded: Whether the statement is logically supported by the context or aligns with indirect evidence. Rule 3: Conclusive: Whether the statement is the only reasonable conclusion supported by indisputable evidence, free from subjective bias.

Table 4: Complete instructions for annotators identifying and evaluating statements in CogniBench.

cination detection:

Tasksource (Sileo, 2024): Tasksource is zero-shot classification nli model, we pair each response sentence against context, and gather all "contradict" sentence as detection output.

SelfCheckGPT (Manakul et al., 2023): Self-CheckGPT implement use fine-tuned DeBERTa-v3-large, it output entailment (or Contradiction) score with input being the sentence and a sampled passage. We pair each response sentence against context, and gather all "contradict" sentence as detection output.

FAVA (Mishra et al., 2024): FAVA (FAct Verification with Augmentation), a model for fine-grained hallucinations detections and editing. FAVA is trained on high-quality synthetic training data to identify hallucinations, incorporating retrieved knowledge. We map spans to sentence-level to ensure a fair comparison.

RAG-Truth (Niu et al., 2023): RAGTruth uses the Llama2 13B instruct fine-tuned with the training set from RAGTruth. The model takes the context-response pair with proper instructions as the input and treats the hallucinate span as the targeted generation output. We mapping spans to sentence-level to ensure a fair comparison.

Auto-Labeling: We employ the auto-labeling pipeline Section 3 on human-labeled CogniBench data to assess the reliability of this method.

CogniDet : We fine-tune the Llama3 8B instruct model (Grattafiori et al., 2024) using the CogniBench-L dataset, which is generated via the auto-labeling pipeline outlined in Section 3.

A.8.2 CogniDet Training Details

We use the CogniBench-L that generated by auto-labeling pipeline proposed in Section 3, to fine-tune Llama3 8B instruct model (Grattafiori et al., 2024). The training setup and hyperparameters are as follows: the epoch is 3, the batch size is 2, the learning rate is 5e-5 We generate responses using sampling implemented via vLLM (Kwon et al., 2023). Our model is trained on 8 NVIDIA A6000 GPUs. It takes approximately 18 hour to train.

CogniDet takes the context-response pair and direct generate a list of hallucinated sentences as response, as shown in Fig. 13.

A.9 Evaluation of Auto-labeling

Hallucination type	Overall		Factual Hallucination		Cognitive Hallucination	
	Recall	Precision	Recall	Precision	Recall	Precision
Auto-Labeling (Threshold = 2)	77.98	87.76	74.75	91.05	78.56	85.55
Auto-Labeling (Threshold = 3)	75.88	89.63	72.72	91.70	76.43	87.83
- Sampling	67.72	88.05	67.98	89.50	66.76	86.33
- CFP	60.49	85.11	53.69	85.26	62.65	84.29

Table 5: Ablation Study of Auto-Labeling Pipeline

We present an ablation study of our proposed

Type of statement	Definition	Examples
Misleading Statement	Fails to satisfy Criterion 1. These statements are incorrect, ambiguous, or deceptive in some way. They may misrepresent facts, make unfounded assumptions, or mislead individuals in the wrong direction.	<p>Context: ...Raymond is a city in Pacific County, Washington, United States. The population was 2,882 at the 2010 census. The 2020 census showed the population of 3081, an increase of 6.4%. The town's economy has traditionally been based on logging and fishing, together with a limited amount of tourism. (800 Words omit)</p> <p>Response: Be sure to spend some time exploring the beautiful natural surroundings, as they contribute significantly to the area's charm. (Annotation reason: Rational? No, makeup belief)</p> <p>Context: Piccadilly Circus station, at the east end of the street, was designed by Charles Holden and built between 1925 – 28. It was the first underground station to have no above-ground premises; the station is only accessible by subways from street level...</p> <p>Response: Holden's design for the station features a modern and minimalist aesthetic, with a focus on functional elements and clean lines. (Annotation reason: Rational? No, the context describes the station's historical construction but does not mention its design aesthetic, likely to mislead user.)</p>
Speculative Statement	Satisfies only Criterion 1. These statements are plausible based on the available information but lack explicit support.	<p>Context: The North Dakota Fighting Hawks (formerly known as the Fighting Sioux) are the athletic teams Having won eight national championships, the men's hockey team is easily the most recognized of UND teams. A charter member of the National Collegiate Hockey Conference....</p> <p>Response: Their (Hawks team) success and reputation attract top talent from across north america and beyond, playing a crucial role in maintaining the team's high standards year after year. (Annotation reason: Rational? Yes, it is rational to infer that success, and reputation may attract talent → Grounded? No, the context does not provide direct evidence that success leads to the attraction of talent.)</p> <p>Context: ...Brandon Dennard Pettigrew (born February 23, 1985) is a former American football tight end. He played college football at Oklahoma State. He was drafted in the first round (20th overall) of the 2009 NFL Draft by the Detroit Lions, and spent his entire 8-year career with the team....</p> <p>Pettigrew was arrested for disorderly conduct and public drunkenness on February 5, 2017, after a Dallas, Texas bar fight.</p> <p>Response: It's important to note that transitioning from a long and successful football career to a life outside the sport can be challenging for many players. (Annotation reason: Rational? Yes, it is generally rational to assume that transitioning from professional sports can be challenging → Grounded? No, this claim is not specifically supported by the context)</p>
Reliable Statement	Satisfies both Criterion 1 and Criterion 2. These statements are grounded in the original text and may include some subjective terms, such as "seriously" or "significantly." While these terms are not precisely measurable, they are contextually relevant and supported by the available information.	<p>Context: ...State Route 516 (SR 516) is a 16.49-mile long (26.54 km) state highway in the U.S. state of Washington, serving communities in southern King County. The highway travels east as the Kent-Des Moines Road and the Kent-Kangley Road from a concurrency with SR 509 in Des Moines through Kent and Covington to an intersection with SR 169 in Maple Valley....</p> <p>Response: ...It (Washington state route 516) serves communities in southern king county, providing an important transportation route in the region... (Annotation reason: Rational? Yes → Grounded? Yes, the context supports the claim that SR 516 serves communities in southern King County → Conclusive? No, the word "important" introduces subjectivity, as what is deemed "important" may vary depending on perspective.)</p> <p>Context: ...In 1937 , radiosondes were used to aide tropical cyclone forecasting. The next decade saw the advent of aircraft-based reconnaissance by the military... The launch of the first weather satellite , TIROS-I , in 1960 , introduced new techniques to tropical cyclone forecasting that remain important to the present day 1960...</p> <p>Response: ...An breakthrough came in 1960 with the launch of the first weather satellite, TIROS-I.... Tropical cyclone track forecasting has evolved significantly over the years, improving in accuracy and methodology. (Annotation reason: Rational? Yes, it is rational to acknowledge a breakthrough in forecasting with the launch of the satellite → Grounded? Yes, the context supports the notion of advancements in forecasting over time → Conclusive? No, subjectivity involved.)</p>
Conclusive Statement	Satisfies all three Criterion. These are statements that are undeniably correct, free from subjective bias, and logically sound. They can be confirmed with absolute certainty, leaving no room for doubt.	<p>Context: ...the vast majority of products within the stores are alcoholic; one must be either over the age of 21 or with an adult over the age of 25 in order to enter the store...</p> <p>Response: This policy helps prevent underage customers from accessing their products. (Annotation process: Rational? Yes, it is rational to infer that the policy helps prevent underage customers → Grounded? Yes, the context mentions age restrictions for entering the store → Conclusive? Yes, the only reasonable interpretation of the policy is that it aims to prevent underage access to products.)</p> <p>Context: Haunting Ground was first revealed in September 2004 Capcom added the dog mechanic during development thinking retailers and players alike would not enjoy a survival horror title with a lone female protagonist... However , the story was found to be cliché , and some of the gameplay was noted as predictable and dated ...</p> <p>Response: Some gameplay aspects were noted as predictable and dated, which had an impact on the overall experience for players (Annotation reason: Rational? Yes → Grounded? Yes; → Conclusive? Yes, dated gameplay inevitable impact the experience.)</p>

Figure 10: Examples of cognitive statements categorized according to the standards outlined in Section 2.2. The red-highlighted text indicates issues with the statement, while the greyed-out text represents the corresponding annotation reasons.

auto-labeling pipeline in Table 5. We evaluate the impact of removing key components: (1) CFP refers to the removal of the Contrastive and Formative Prompting (CFP) approach as described in Section 3.1, and (2) Sampling indicates the exclusion of the Multi-response Sampling method introduced in Section 3.2. The results highlight the importance of both techniques, as their absence leads to a noticeable drop in recall and precision, especially for cognitive hallucinations. For Auto-Labeling, we apply five-time sampling, where each sentence is annotated five times with identical prompts, and a majority-vote criterion is applied to determine its final classification. This offers flexibility in adjusting the trade-off between precision and recall. The threshold refers to the decision boundary for classifying a sentence as faithful or hallucinated: a higher threshold increases precision by demand-

ing stricter agreement across the responses, while a lower threshold enhances recall by relaxing the agreement requirement.

A.10 Example of Hallucination Detection of CogniDet

We present a hallucination detection example in Fig. 13. CogniDet takes Context and Dialogue as input and outputs a list of hallucinated statements, including invented, speculative, and misleading statements, all detected in a single forward pass.

A.11 Performance of CogniDet versus training data size

Fig. 12 reveals a clear log-linear relationship between training data size and model performance, suggesting continued performance improvements with additional data. This scaling law demonstrates



Figure 11: CogniBench employs a uniform sampling across various topics to ensure conversational diversity. The size of each bubble represents the proportion of that topic in the benchmark

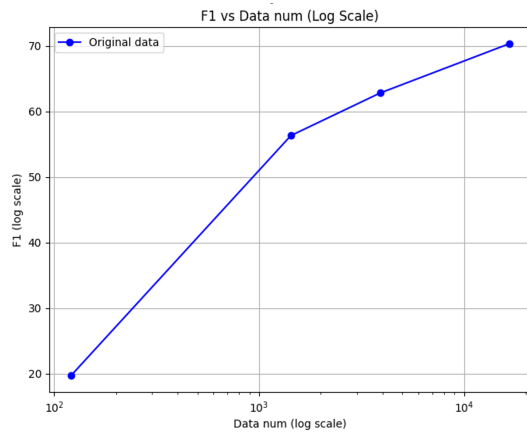


Figure 12: CogniDet’s detection accuracy (F1) showing log-linear scaling with training data size

the importance of large-scale annotated datasets like CogniBench-L for developing robust hallucination detection systems.

A.12 Compare between CogniBench and existing benchmarks

We compare the query context and response from RAGTruth (Niu et al., 2023) and CNN/Daily Mail (See et al., 2017) in Fig. 14. Both RAGTruth and CNN/Daily Mail dataset are composed mainly with factual statement. RAGTruth answers focus more on rephrase the information from the retrieved passages. CNN/Daily Mail excels in news summarization, where the response retrieval key factual information from context to generate a con-

densing article.

CogniBench differentiates itself from these previous datasets by not only emphasizing factual recall but also generating cognitive statements—such as inferences, evaluations, and opinions. For example, in the case of *Moringa oleifera* cultivation (shown in Fig. 14), the generated response includes not only factual details about its growing conditions but also reasoning about how climate and soil factors can influence flowering patterns. This type of response moves beyond simple fact recall, offering more complex insights that align with cognitive processes like evaluation and inference.

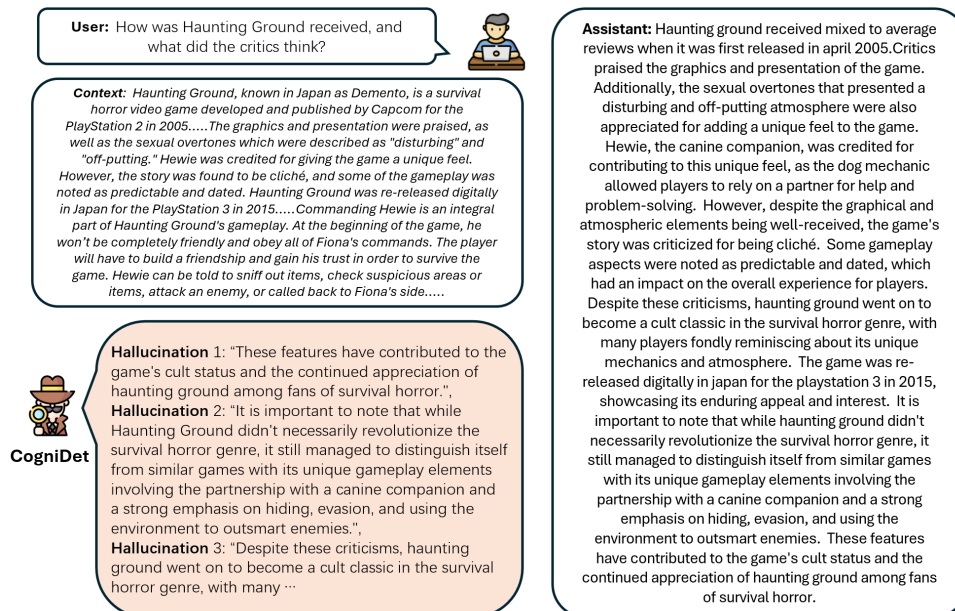


Figure 13: Example of hallucination detection output from CogniDet, trained on CogniBench-L

Dataset	Query	Context	Response
Cogni-bench	I want to plant a moringa oleifera tree in my garden, but i live in a region with cool seasonal temperatures. can you help me understand whether it will flower, and if so, how often?	<p>Moringa oleifera is a fast-growing, drought-resistant tree of the family Moringaceae, native to the Indian subcontinent.</p> <p>....</p> <p>Flowering begins within the first six months after planting. In seasonally cool regions, flowering only occurs once a year in late Spring and early Summer (northern hemisphere between April and June, southern hemisphere between October and December). In more constant seasonal temperatures and with constant rainfall, flowering can happen twice or even all year-round.</p> <p>....</p> <p>The moringa tree is grown mainly in semiarid, tropical, and subtropical areas, corresponding in the United States to USDA hardiness zones 9 and 10. It tolerates a wide range of soil conditions, but prefers a neutral to slightly acidic (pH 6.3 to 7.0), well-drained, sandy or loamy soil. In waterlogged soil, the roots have a tendency to rot. Moringa is a sun- and heat-loving plant, and does not tolerate freezing or frost. Moringa is particularly suitable for dry regions, as it can be grown using rainwater without expensive irrigation techniques.</p> <p>....</p>	<p>Moringa oleifera is primarily suited for semiarid, tropical, and subtropical regions, corresponding to usda hardiness zones 9 and 10.</p> <p>It is a sun- and heat-loving plant that does not tolerate frost or freezing temperatures well.</p> <p><u>In areas with cool seasonal temperatures or frost, the growth and flowering of the plant may be limited or impacted.</u></p> <p>In regions with more constant seasonal temperatures and regular rainfall, moringa oleifera can flower twice or even all year round.</p> <p>However, in seasonally cool regions, flowering generally occurs once a year, late in spring to early summer.</p> <p>For the northern hemisphere, this would be between april and june, and for the southern hemisphere, between october and december.</p> <p><u>If you decide to plant a moringa oleifera tree in your garden, it's essential to ensure that it's protected from freezing temperatures or frost.</u></p> <p><u>You may also consider growing it in a large container that can be moved indoors during the colder months to maintain favorable conditions.</u></p> <p><u>Remember that moringa oleifera prefers a neutral to slightly acidic, well-drained sandy or loamy soil for optimal growth.</u></p> <p><u>It's important to acknowledge that the tree's flowering and fruiting might be limited or restricted when planted in a region with cool seasonal temperatures.</u></p> <p><u>Nonetheless, experimenting with its cultivation and providing a suitable environment could yield a plant capable of flowering and producing pods.</u></p>
		<p>passage 1:1 Dump a chimney of lit coals into the grill. 2 With a grill hoe or garden hoe, spread the coals into a double layer at the far side and a single layer in the center. 3 Leave the section closest to you coal-free. 4 Replace the grill grate. 5 If using a gas grill, set it up for three-zone grilling. passage 2:1 Leave the section closest to you coal-free. 2 Replace the grill grate. 3 If using a gas grill, set it up for three-zone grilling. 4 Keep it hot, keep it clean, keep it lubricated: When the grill grate is hot, clean it with a grill brush, then lubricate it with a grill oiler or a folded paper towel dipped in vegetable oil. passage 3:# 4. The Porterhouse is a man\u2019s steak if there ever was one. It\u2019s basically two steaks in one the NY Strip on one side of the bone and the filet on the other. The bone gives both cuts tremendous flavor and when cooked just right, you won\u2019t find a better steak.</p>	<p>Based on the given passages, here are the steps to grill a Porterhouse:</p> <ol style="list-style-type: none"> 1. Dump a chimney of lit coals into the grill (Passage 1). 2. Spread the coals into a double layer at the far side and a single layer in the center (Passage 1). 3. Leave the section closest to you coal-free (Passages 1 and 2). 4. Replace the grill grate (Passages 1 and 2). 5. Set up the grill for three-zone grilling if using a gas grill (Passage 2). 6. Keep the grill grate hot, clean, and lubricated (Passage 2). 7. Cook the Porterhouse until it reaches your desired level of doneness.
RAGTruth	how to grill a porterhouse		
CNN / Daily Mail	Summary	<p>Harry Potter star Daniel Radcliffe gains access to a reported £20 million (\$41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him.</p> <p>....</p> <p>To the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties.</p> <p>....</p> <p>Radcliffe's earnings from the first five Potter films have been held in a trust fund which he has not been able to touch.</p> <p>....</p>	<p>Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday .</p> <p>Young actor says he has no plans to fritter his cash away .</p> <p>Radcliffe's earnings from first five Potter films have been held in trust fund</p>

Figure 14: Example dialoge from CogniBench with RAGTruth (Niu et al., 2023) and CNN/Daily Mail (See et al., 2017). Cognitive statements are bold and dotted underlined

SYSTEM PROMPT

Your task is to evaluate the faithfulness of claims. Tag the statements enclosed in the empty tag spaces provided by the user.

Tags to Use:

<faithful>: The sentence is fully verified by the references of the conversation.

<hallu>: hallucination, use this tag for statements that reflect unreliable personal opinions or judgments or statements that include information not found in the references or context

<irrelevant>: The sentence is irrelevant to the reference or does not provide any useful information.

USER PROMPT

Example of incorrect annotation:**Reference:**

Jacques Michel Gabriel Paul Benoist-Méchin (1 July 1901 – 24 February 1983) was a French far right politician and writer...

Claims:

This claim should be tagged as hallu: "In 1936, Jacques benoist-méchin joined the french popular party..." since french popular party is far-right political group opposed to democracy is not mentioned in the reference even though Jacques benoist-méchin was indeed a far right politician.

Example of incorrect annotation:

Reference: The title-safe area or graphics-safe area is, in television broadcasting, a rectangular area which is far enough in from the four edges, such that text or graphics show neatly...

Claims:

This claim should be tagged as faithful: "Content placed in Title-safe area remains legible and undistorted..." since the claim can be reasonably inferred from the reference.

You SHOULD refer to the following correct examples:**Reference:**

Izumi Station is a railway station on the Jōban Line in the city of Iwaki, Fukushima, Japan, operated by East Japan Railway Company (JR East). The station also has a freight depot for the Fukushima Rinkai Railway Main Line.

Claims:

<>Izumi station has a combination of one island platform and one side platform, which are connected to the station building by a footbridge.</> <>The island platform allows trains to pass on either side, while the side platform is used for trains arriving from just one direction.</> <>This layout ensures efficient handling of the passenger and freight trains that pass through the station.</>

Annotation:

<faithful>Izumi station has a combination of one island platform and one side platform, which are connected to the station building by a footbridge.</faithful> <hallu>The island platform allows trains to pass on either side, while the side platform is used for trains arriving from just one direction.</hallu> <hallu>This layout ensures efficient handling of the passenger and freight trains that pass through the station.</hallu>

Now, based on the examples above, please evaluate the following claim:

Reference: {reference}

Claims: {dialogue}

Table 6: System and user prompts for assessing the faithfulness of claims.