# HUDEX: INTEGRATING HALLUCINATION DETECTION AND EXPLAINABILITY FOR ENHANCING THE RELIABILITY OF LLM RESPONSES

# **Sujeong Lee**

Inha University
Incheon, 22212, Republic of Korea
tnwjd025611@inha.edu

# **Hayoung Lee**

Inha University
Incheon, 22212, Republic of Korea
gkdud000123@gmail.com

#### Seongsoo Heo

Inha University Incheon, 22212, Republic of Korea woo555813@inha.edu

#### Wonik Choi

Inha University
Incheon, 22212, Republic of Korea
wichoi@inha.ac.kr

February 13, 2025

#### **ABSTRACT**

Recent advances in large language models (LLMs) have shown promising improvements, often surpassing existing methods across a wide range of downstream tasks in natural language processing. However, these models still face challenges, which may hinder their practical applicability. For example, the phenomenon of hallucination is known to compromise the reliability of LLMs, especially in fields that demand high factual precision. Current benchmarks primarily focus on hallucination detection and factuality evaluation but do not extend beyond identification. This paper proposes an explanation enhanced hallucination-detection model, coined as HuDEx, aimed at enhancing the reliability of LLM-generated responses by both detecting hallucinations and providing detailed explanations. The proposed model provides a novel approach to integrate detection with explanations, and enable both users and the LLM itself to understand and reduce errors. Our measurement results demonstrate that the proposed model surpasses larger LLMs, such as Llama3 70B and GPT-4, in hallucination detection accuracy, while maintaining reliable explanations. Furthermore, the proposed model performs well in both zero-shot and other test environments, showcasing its adaptability across diverse benchmark datasets. The proposed approach further enhances the hallucination detection research by introducing a novel approach to integrating interpretability with hallucination detection, which further enhances the performance and reliability of evaluating hallucinations in language models.

# 1 Introduction

Recent advancements in large language models (LLMs) have showcased their potential in natural language processing (NLP) [1]. While LLMs can generate effective responses across diverse tasks, they are also limited by certain critical issues. One such limitation is hallucination, where the model produces information that is factually incorrect or generates content not requested or instructed by the user. This problem can lead to the spread of incorrect information, particularly problematic in fields where accuracy and reliability are crucial, thereby limiting the applicability of LLMs in various industries. Consequently, hallucination is a major issue undermining the reliability of LLMs, prompting significant research into solutions.

Recent studies have focused on developing benchmarks to detect and evaluate hallucinations and methods for mitigating them. For example, FELM [2] provides a benchmark for assessing the factuality of LLMs by identifying factual errors in response segments through text-segment-based annotations. TruthfulQA [3] evaluates whether language models

produce truthful responses, aiming to detect non-factual responses across various domains. Similarly, QAFactEval [4] proposes a QA-based metric for assessing factual consistency in summarization tasks, effectively detecting and evaluating factual errors.

However, these studies primarily focus on evaluating or detecting hallucinations or a lack of factual inaccuracies, rather than actively improving the model's reliability. This limitation underscores the need for approaches that not only assess factual errors but also actively contribute to improving the quality of model responses. Additionally, benchmark-based evaluation methods may struggle with the real-time detection of hallucinations in model-generated responses.

To address these gaps, we propose a specialized model named HuDex designed to detect hallucinations in LLM responses and provide detailed explanations of these hallucinations. Unlike existing benchmarks, our model not only identifies hallucinations but also offers specific explanations, helping users understand the model's output and assisting the model in refining its responses. This approach aims to improve the reliability of LLM responses.

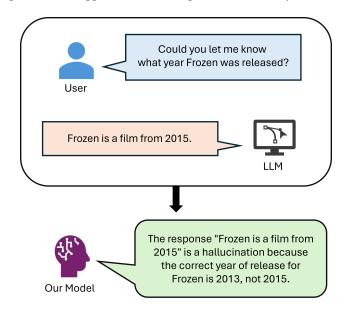


Figure 1: Schematic Representation of Our Hallucination Model

The key contributions of our proposed model are:

- 1. Moving beyond standardized hallucination benchmarks, the proposed model enables proactive detection despite its smaller size.
- 2. By providing detailed explanations of detected hallucinations, the model enhances user understanding and contributes to the improvement of model performance.
- 3. Through an analysis focused on hallucinations, a more nuanced evaluation of the hallucination domain is possible compared to general-purpose LLMs, and this can be effectively used to evaluate other LLMs.

## 2 Related Work

#### 2.1 Definitions of Large Language Models

A Large Language Model is an artificial intelligence model based on the Transformer architecture [5]. It refers to a pre-trained language model (PLM) with a parameter size exceeding a certain threshold [6]. LLMs are trained on massive datasets and typically have billions to hundreds of billions of parameters. Due to the extensive data used in their training, LLMs exhibit exceptional performance across various NLP tasks, including text generation, translation, and summarization.

Notably, LLMs that surpass a certain parameter scale demonstrate emergent abilities not found in smaller models. Examples of these abilities include in-context learning, instruction following, and chain-of-thought (CoT) reasoning [7]. These capabilities enable LLMs to handle more complex tasks, such as advanced reasoning, problem-solving, and generating multi-turn responses.

Although LLMs are primarily used for general downstream tasks, their increasing significance in both academia and industry has led to research into domain-specific LLMs. Examples include the Med-PaLM series for the medical domain [8] and FinGPT for the financial domain [9]. These advancements underscore the growing need for LLMs not only in language generation but also in addressing specialized tasks across various fields.

#### 2.2 Definitions of Hallucination

In NLP, hallucination refers to content that is unreliable or illogical compared to the provided source material [10], [11]. Previous studies categorize hallucinations into two broad two types: intrinsic and extrinsic [10], [11], [12], [13].

Intrinsic hallucination occurs when the generated output contradicts the source content. For example, this happens when a model produces information that conflicts with the given data in response to a factual question. In contrast, extrinsic hallucinations involve outputs that include unverifiable or nonexistent information. This often occurs when the model generates content that cannot be corroborated by the source material.

In the context of LLMs, hallucination can be defined more specifically. LLM hallucinations, which prioritize user instructions and interactions, can be categorized based on factuality and faithfulness [14]. Factual hallucinations arise when a model generates outputs that are based on real-world information but are either incorrect or unverifiable. For instance, if the model inaccurately presents well-known facts or mentions nonexistent information, it is considered a factual hallucination. Faithfulness-related hallucinations occur when the model generates responses unrelated to user instructions or the provided content, or when it produces internally inconsistent answers. This type of hallucination is particularly important in conversational models.

The issue of hallucination may stem from several factors, including the use of outdated data during the data collection process [15] or biased data [16] used for model training [14],[17], [18]. Furthermore, the risk of hallucinations tends to increase with the size and complexity of the models.

#### 2.3 LLM-Based Evaluation of LLMs

One of the key challenges discussed alongside the development of LLMs is the difficulty in accurately evaluating the context and meaning of generated responses using traditional quantitative metrics. While human evaluation has been employed to address this limitation, it has considerable drawbacks, particularly in terms of time and resource consumption [1],[19].

To overcome these challenges, the use of LLMs as evaluation tools, or "LLM judges," has gained attention. [20] pioneered an LLM-based evaluation framework, showing that strong LLMs achieved over 80% agreement with human experts in evaluations. Subsequent studies by [21], [22], and [23] have expanded on this approach, further validating the utility of LLM judges.

The introduction of LLM judges provides an efficient solution for evaluating large-scale data, where human evaluation may be impractical. In addition to quantitative assessments, LLM judges offer qualitative evaluations based on their understanding of context and adherence to user instructions, making them versatile tools for comprehensive evaluation.

#### 3 Data Construction

## 3.1 Datasets

For training, we utilized the HaluEval, FactCHD, and FaithDial datasets, as summarized in Table 1.

The HaluEval dataset [24] is a hallucination evaluation benchmark designed to assess the likelihood of hallucinations based on content type. It consists of 30,000 examples across three tasks: question answering, knowledge-based dialogues, and text summarization, along with 5,000 general user queries that include ChatGPT responses. In this study, we used the question-answering and knowledge-based dialogue subsets as training data. Both subsets focus on detecting hallucinations based on provided knowledge, allowing the model to learn how to identify intrinsic hallucinations.

The FactCHD dataset [25] is a benchmark specifically designed to detect hallucinations that conflict with factual information in LLMs. It evaluates factual accuracy in the context of a wide range of queries and responses, facilitating factual reasoning during evaluation. Unlike HaluEval, the FactCHD dataset does not include a pre-existing knowledge base, enabling the model to learn to detect hallucinations in scenarios with limited reference material.

The FaithDial dataset [26] is designed to minimize hallucinations and improve the accuracy of information-seeking dialogues. It was built by modifying the Wizard of Wikipedia (WOW) benchmark to include hallucinated responses. The dataset includes a BEGIN [27] label that categorizes responses based on their relationship to the knowledge source

and their contribution to the dialogue. For binary classification of hallucination detection, we preprocessed the dataset by excluding the Generic and Uncooperative categories. Additionally, since each data point includes both a response and an original response, we split them into two distinct responses. This allowed us to create two separate data instances with the same knowledge and dialogue context but different responses, thereby augmenting the training data.

Table 1: Dataset Information

Dataset	Train	Test
HaluEval Dialogue	9,000	1,000
HaluEval QA	9,000	1,000
FaithDial	18,357	3,539
FactCHD	51,838	6,960

# 3.2 Explanation Generation

The primary goal of our model is not only to detect hallucinations in generated responses but also to provide explanations for the reasoning behind these judgments. A simple example of this process is illustrated in Figure 1. To achieve this, the model must be trained on explanation data. While the FactCHD dataset includes explanations, the HaluEval and FaithDial datasets do not. Therefore, we used the Llama3 70B [28] model to generate explanation data for hallucination detection in the HaluEval and FaithDial datasets.

During the explanation generation process, we also generated answers corresponding to the hallucination labels. This step ensured that the hallucination labels predicted by the model during explanation generation aligned with the existing hallucination labels in HaluEval and FaithDial datasets.

Upon analyzing the model's predictions, we found that 0.5% of the responses failed to understand the prompt and asked for clarification, and 4.2% were classified as anomalies. Excluding these cases, 95.3% of the responses adhered to the expected format. As shown in Table 2, the accuracy of valid responses was 98.3%. Ultimately, 93.7% of the hallucination labels from HaluEval and FaithDial matched the model's predicted answers, and only the verified matching data were used for training.

To further assess the quality of the generated explanations, we conducted statistical sampling. We defined the population as the set of generated explanations, with a confidence level of 99%, a conservatively estimated defect rate of p = 0.5, and a margin of error set at 2%. Through human evaluation of the selected sample, we validated the explanations to ensure the accuracy and relevance of the reasoning provided.

Table 2: Confusion Matrix of Model Predictions VS Actual Answers (Proportional Representation)

	ActualPositive	ActualNegative
Predicted Positive	52.0%	1.7%
Predicted Negative	0%	46.3%

# 4 Model Training and Inference

# 4.1 Training

We used the Llama 3.1 8B model [28] for training and applied low-rank adaptation (LoRA) [29], a method under parameter efficient tuning (PEFT). The task prompts for training were divided into two main categories: hallucination detection and hallucination explanation. The model was trained on both tasks using the same dataset.

#### 4.2 Inference

The prompt structure for inference focuses on two key elements: persona provision and task stage provision. Persona provision ensures that the model understands the specific task's goal before generating responses, encouraging deeper analysis of the given information. By defining the task's context and role in advance, we aim for more consistent outputs. To generate a persona, we provided ChatGPT with task details and received recommendations for suitable persona candidates. After a human filtering process, we selected a hallucination expert persona to detect hallucinations.

Task stage provision guides the model to approach complex problems systematically when generating responses. The prompt stages are structured adaptively based on the task and data characteristics. If background knowledge is available,

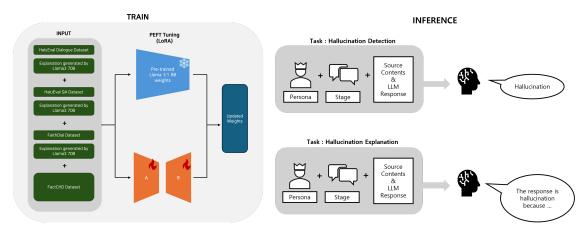


Figure 2: Overview of HuDEx: Training and Inference

the model generates responses based on it; otherwise, it relies on context and its inherent knowledge. The stage structure also varies depending on whether the task focuses on hallucination detection or explanation generation. Stages can be divided or combined based on the specific needs of each task.

An overview of the training and inference process can be found in Figure 2, and brief examples of both the stage and persona structures are shown in Figure 3.

# 5 Experiments

#### 5.1 Datasets

For the detection and explanation generation experiments, we used the test sets from HaluEval dialogue, HaluEval QA, FaithDial and FactCHD, which were also used during training. The HaluEval datasets, both for dialogue and QA tasks, provide background knowledge, so we applied inference prompts designed to incorporate this information. FaithDial also utilized the same inference prompt. For the FactCHD dataset, which does not include background knowledge, we used the inference prompt stages suited for tasks without background knowledge. The persona was consistently provided across all tasks, regardless of the presence or absence of background knowledge.

For zero-shot detection, we conducted experiments on HaluEval subsets not used during training, specifically HaluEval summarization and HaluEval general. The HaluEval summarization dataset focuses on detecting hallucinations in document summaries, while the HaluEval general dataset evaluates hallucination detection in ChatGPT responses to user queries. Since both datasets lack background knowledge, we used inference prompts designed for tasks without background knowledge.

## 5.2 Test Setting

#### **5.2.1** Detection Experiments

For the detection experiments, we compared our HuDEx to two LLMs, GPT-4 [30] and Llama 3 70B. These models received the same inference prompts as our model and were tasked with classifying whether the responses contained hallucinations.

# **5.2.2** Explanation Generation Experiments

To evaluate the explanations generated by each model, we used an LLM judge and conducted main experiment. The experiment followed a single-answer grading approach, where each model's response was individually scored.

In the single-answer grading experiment, we divided the evaluation into two categories: factuality and clarity. Factuality assessed whether the explanation contained hallucinations, contradictions, or accurately reflected the given information. Clarity evaluated how clearly and thoroughly the reason was articulated. Each criterion was scored on a 3-point scale, with a maximum total score of 6 points.



- Expert in detecting Al-generated hallucinations (incorrect or unsupported information)
- Ph.D. in Computational Linguistics
- · Over 15 years of experience in Natural Language Processing (NLP) and AI ethics
- Primary task: Identify hallucinations in AI responses based on provided context





- 1. Review background knowledge to identify relevant facts
- 2. Identify key elements in the content's main points or flow
- 3. Compare the response with background facts, conversation flow, or user request
- 4. Evaluate if the response aligns with the facts and request or contains unsupported information
- 5. Determine hallucination and provide an explanation for the decision





- 1. Identify key elements in the conversation flow
- 2. Compare the response with the conversation flow and user request
- 3. Evaluate if the response aligns with the request and maintains logical consistency
- 4. Determine hallucination and explain the decision

Figure 3: Examples of Persona and Steps Used in Inference Prompts

We used GPT40 as the judge for experiment. In the HaluEval and FaithDial dataset, we compared the explanations generated by our model against those from Llama 3 70B, with GPT40 providing the final judgments. For the FactCHD dataset, we compared the explanations generated by HuDEx against the explanations included in the FactCHD dataset itself.

#### 6 Results

#### 6.1 Detection Results

# 6.1.1 Test Data Detection

In this experiment, binary classification was performed to distinguish hallucinations from non-hallucinations using the test sets from the training data, with accuracy as the evaluation metric. Table 3 compares the performance of Llama3 70B, GPT40, and our model across benchmark datasets such as HaluEval dialogue, HaluEval QA, FactCHD, and FaithDial.

The experimental results show that our HuDEx outperformed the larger models, Llama3 70B and GPT4o, across all benchmarks. Specifically, it achieved an accuracy of 80.6% on the HaluEval dialogue dataset, surpassing Llama3 70B (71.8%) and GPT4o (72.5%), indicating superior performance in detecting hallucinations in conversational response.

In the HaluEval QA dataset, our model again achieved the highest accuracy of 89.6%, outperforming GPT4o (86.6%) and Llama3 70B (82.7%). This demonstrates its refined ability to detect hallucinations in QA tasks.

On the FactCHD and FaithDial datasets, HuDEx recorded accuracies of 70.3% and 58.8%, respectively, continuing to show strong performance on both benchmarks. On the FactCHD dataset, HuDEx outperformed Llama3 70B by 11%, confirming its effectiveness in hallucination detection even when background knowledge is unavailable. On the FaithDial dataset, our HuDEx also significantly outperformed GPT4o (50.6%), achieving 58.8%, which highlights its consistent performance on a different type of conversation-based dataset compared to HaluEval dialogue.

These results demonstrate that our model consistently delivers superior performance in hallucination detection across various benchmark datasets, outperforming larger models.

Table 3: Test Data Detection Results (Accuracy)

Model	HaluEval Dialogue	HaluEval QA	FactCHD	FaithDial
Llama3 70B	71.8	82.7	59.4	47.9
GPT4o	72.5	86.6	61.2	50.6
HuDEx	80.6	89.6	70.3	58.8

#### 6.1.2 Zero-Shot Detection

Table 4 presents the results of the binary classification experiment on hallucination vs. non-hallucination in a zero-shot setting. This experiment evaluated the model's hallucination detection performance on unseen data using the HaluEval summarization and HaluEval general datasets, which were not included in the training data. Accuracy was used as the evaluation metric, consistent with the methodology in previous experiments.

On the HaluEval summarization dataset, HuDEx achieved an accuracy of 77.9%, outperforming Llama3 70B (69.55%) and GPT4o (61.9%). This demonstrates the model's ability to effectively detect hallucinations in summary texts of original content.

The HaluEval general dataset consists of queries posed by real users to GPT models, often containing complex responses that go beyond typical conversational text. This complexity makes hallucination detection more challenging and serves as an important benchmark for evaluating model reliability on unstructured data. On this dataset, GPT4o recorded the highest accuracy at 78.0%, while our model achieved 72.6%. These results suggest that while HuDEx delivers consistent performance on complex responses, there is still room for improvement.

Table 4: Zero-shot data detection results (Accuracy)

Model	HaluEvalSummarization	HaluEvalGeneral
Llama3 70B	69.55	76.2
GPT4o	61.9	78.0
HuDEx	77.9	72.6

## **6.2** Explanation Generation Results

# 6.2.1 Single-Answer Grading

This experiment presents the evaluation of hallucination explanations generated by Llama3 70B and our model, as assessed by the LLM judge. The results, shown in Table 5, were obtained from the HaluEval dialogue, HaluEval QA, and FaithDial datasets. Explanations were evaluated based on two criteria: factuality and clarity, each scored out of 3 points, for a maximum combined score of 6 points.

When comparing the performance of Llama3 70B and our HuDEx in terms of factuality, Llama3 70B scored lower on the HaluEval dialogue dataset with 1.932 points but achieved relatively higher scores on HaluEval QA and FaithDial, with 2.416 and 2.587 points, respectively. In contrast, our model outperformed Llama3 70B on factuality for the HaluEval dialogue dataset, though it scored slightly lower on HaluEval QA (2.299) and FaithDial (2.216). Despite the variations in scores across datasets, HuDEx demonstrated consistent factual accuracy, indicating its ability to provide reliable information.

In terms of clarity, Llama3 70B achieved the highest score on the FaithDial dataset with 2.451 points, while our model closely followed with 2.417 points. On the HaluEval dialogue and HaluEval QA datasets, our model outperformed Llama3 70B, scoring 2.413 and 2.523 points, respectively. This indicates that HuDEx provides clearer and more easily understandable explanations for hallucinations.

Overall, our HuDEx demonstrated competitive performance in terms of factuality, clarity, and overall scores compared to Llama 70B. These results support that our model consistently delivers reliable and clear hallucination explanations.

The next experiment evaluated the original explanations from the FactCHD dataset against those generated by our model, with results shown in Table 6. The conversion ratio was used to compare the performance of our HuDEx as a percentage, with the FactCHD score serving as the maximum (100%).

For factuality, FactCHD recorded a score of 2.2549, while our model scored slightly lower at 2.236. The conversion ratio for factuality was 99%, indicating that although FactCHD's original explanations had slightly higher factual accuracy, HuDEx performed very closely to this benchmark.

Table 5: Comparison of Hallucination Explanations Between Llama 370B and Proposed model (LLM Judge Evaluation)

Model	Dataset	Factuality (3)	Clarity (3)	Overall (6)
Llama3 70B	HaluEval Dialogue	1.932	2.302	4.256
	HaluEval QA	2.416	2.153	4.569
	FaithDial	2.587	2.451	5.038
HuDEx	HaluEval Dialogue	2.116	2.413	4.528
	HaluEval QA	2.299	2.523	4.822
	FaithDial	2.216	2.417	4.633

In terms of clarity, FactCHD achieved a score of 2.439, while our model scored slightly lower at 2.37. The conversion ratio for clarity was 97%, suggesting that while our model's explanations were marginally less clear than FactCHD's, they remained highly comparable in clarity. In conclusion, HuDEx showed performance similar to FactCHD, with conversion ratios ranging from 97% to 99%. These results demonstrate that HuDEx generates explanations nearly equivalent in quality to the original explanations provided in the FactCHD dataset.

Table 6: LLM Judge Evaluation of Explanations:FactCHD original vs HuDEx

	Factuality (3)	Clarity (3)	Overall (6)
FactCHD	2.2549	2.439	4.697
HuDEx	2.236	2.37	4.61
Conversion Ratio	99%	97%	98%

## 7 Conclusion

The hallucination phenomenon in large language models (LLMs) presents a significant challenge that needs to be addressed in practical applications. This study proposes a model called HuDEx specifically designed to detect hallucinations in LLM-generated responses and provide explanations for them. By offering such feedback, the model contributes to both user understanding and the improvement of LLM, fostering the generation and evaluation of more reliable responses.

However, a key limitation of the model is its reliance on the LLM's inherent knowledge when sufficient source content is unavailable for detecting and explaining hallucinations. This dependency can reduce the clarity of the explanations and, in some cases, introduce hallucinations into the explanations themselves.

Despite this limitation, the study demonstrates strong potential for detecting and explaining hallucinations. Future research should focus on overcoming these challenges and exploring methods to improve the model's performance. For example, integrating external knowledge retrieval systems could reduce the model's reliance on its internal knowledge, while enhancing reasoning-based validation could lead to more reliable explanations.

Additionally, we aim to develop an automated feedback loop in future work. This system would allow for continuous correction and improvement of hallucinations, contributing to greater reliability and consistency in LLMs over time.

# References

- [1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [2] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. Felm: Benchmarking factuality evaluation of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- [3] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. Association for Computational Linguistics, 5 2022.
- [4] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza

- Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601. Association for Computational Linguistics, 7 2022.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł{}ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc., 2017.
- [6] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.
- [7] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.
- [8] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172–180, 8 2023.
- [9] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*, 2023.
- [10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55, 3 2023.
- [11] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics, 7 2020.
- [12] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214. Association for Computational Linguistics, 11 2021.
- [13] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey, 2023.
- [14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.
- [15] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. Realtime qa: what's the answer right now? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024.
- [16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT 2021 Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery, Inc, 3 2021.
- [17] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention, 2022.
- [18] Zuchao Li, Shitou Zhang, Hai Zhao, Yifei Yang, and Dongjie Yang. Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer, 2023.
- [19] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In Jong C Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718. Association for Computational Linguistics, 11 2023.

- [20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mtbench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024.
- [21] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. Benchmarking foundation models with language-model-as-an-examiner, 2023.
- [22] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics, 12 2023.
- [23] Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations, 2023.
- [24] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464. Association for Computational Linguistics, 12 2023.
- [25] Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. Factchd: Benchmarking fact-conflicting hallucination detection, 2024.
- [26] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022.
- [27] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083, 2022.
- [28] Llama Team and Ai @ Meta. The llama 3 herd of models, 2024.
- [29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representa*tions, 2022.
- [30] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reijchiro Nakano, Rajeey Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul

Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.