ArxEval: Evaluating Retrieval and Generation in Language Models for Scientific Literature

Anonymous authors
Paper under double-blind review

Abstract

Language Models [LMs] are now playing an increasingly large role in information generation and synthesis; the representation of scientific knowledge in these systems needs to be highly accurate. A prime challenge is hallucination; that is, generating apparently plausible but actually false information, including invented citations and nonexistent research papers. This kind of inaccuracy is dangerous in all the domains that require high levels of factual correctness, such as academia and education. This work presents a pipeline for evaluating the frequency with which language models hallucinate in generating responses in the scientific literature. We propose **ArxEval**, an evaluation pipeline with two tasks using ArXiv as a repository: **Jumbled Titles** and **Mixed Titles**. Our evaluation includes fifteen widely used language models and provides comparative insights into their reliability in handling scientific literature.

1 Introduction

Large Language Models (LLMs) have emerged as pivotal tools in information access and generation, particularly through their capabilities of producing factually accurate texts. As these models become increasingly integrated into various applications, ensuring the accuracy of their responses has become very important. The performance and reliability of LLMs in generating accurate information are significantly influenced by multiple factors, including training data quality, model architecture design, and post-training optimization processes Naveed et al. (2024), Minaee et al. (2024), Guo et al. (2023).

However, a significant challenge in the deployment of LLMs lies in their propensity to generate nonfactual responses, a phenomenon commonly referred to as hallucination. These hallucinations fundamentally undermine the reliability and faithfulness of LLMs, presenting substantial obstacles to their widespread adoption across various domains Huang et al. (2024), Sahoo et al. (2024). The mitigation of hallucinations has consequently emerged as a critical area of research within the field. While various strategies have been proposed and implemented to reduce hallucinations, showing promising improvements in the faithfulness of LLMs for general-purpose tasks, domain-specific applications remain particularly challenging Tonmoy et al. (2024), Rawte et al. (2023), Berberette et al. (2024).

In this paper, we present a comprehensive study evaluating the extent of hallucination in LLMs under domain-specific prompting, with a particular focus on scientific literature. We develop and implement a systematic evaluation pipeline to assess fifteen prominent open-source LLMs: Qwen 2.5 Yang et al. (2024), Gemma 2 Team et al. (2024), Llama 3 Grattafiori et al. (2024), Phi 3 Abdin et al. (2024), Orca 2 Mitra et al. (2023), Mistral v-0.3 [Team (2024), Deepseek-llm DeepSeek-AI et al. (2024), Olmo-2 OLMo et al. (2024), Mistral-Nemo Team, Eurus-2 Yuan et al. (2024), and Solar-Pro upstage (2024). Our evaluation utilizes the ArXiv dataset Clement et al. (2019) as the primary source of scientific articles, providing a robust foundation for assessing model performance in academic contexts.

The evaluation pipeline **ArxEval** introduces two novel tasks: **Jumbled Titles** and **Mixed Titles**. These tasks are specifically designed to assess the faithfulness of LLMs in retrieving and reasoning about scientific articles under challenging conditions. The models are presented with either jumbled or mixed titles and evaluated not only on their prompt adherence but also on the quality and accuracy of their generated outputs. By adopting an open-ended evaluation approach, we aim to provide comprehensive insights into the models' capabilities in processing and responding to ambiguous or altered inputs within a domain-specific context, particularly focusing on their ability to maintain factual accuracy while handling complex scientific information.

This study contributes to the growing body of research on LLM reliability and provides valuable insights into the current limitations and capabilities of state-of-the-art language models in handling domain-specific tasks. Our findings have important implications for the development and deployment of LLMs in scientific and academic applications, where maintaining factual accuracy is crucial.

2 Related Work

2.1 Hallucinations in Large Language Models (LLMs)

Hallucinations in LLMs have been extensively studied and documented. While significant advancements have been made in improving their accuracy and reliability, LLMs have been shown to hallucinate even when tasked with generating responses based on known facts Jiang et al. (2024). Such behavior suggests an inherent limitation in these models, reinforcing the hypothesis that hallucination may be an intrinsic characteristic Banerjee et al. (2024).

2.2 Hallucinations in Domain-Specific Settings

2.2.1 Definition and Challenges

Domain-specific hallucinations manifest when LLMs generate inaccurate or fabricated information in specialized fields. In domains like biomedicine, such hallucinations can have serious implications, potentially leading to incorrect medical advice or misinterpretation of research data. The fundamental challenge lies in maintaining factual accuracy while preserving the model's ability to generate coherent and contextually relevant responses.

2.2.2 Causes and Perspectives

Domain-specific hallucinations primarily stem from two factors: deficiencies in training data and limitations in model architecture Dziri et al. (2022). However, recent research presents an alternative viewpoint, suggesting that under certain conditions, hallucinations could be leveraged as a resource for novel problem-solving approaches Sui et al. (2024).

2.2.3 Detection and Evaluation Frameworks

- **DelucionQA** Sadat et al. (2023): A specialized dataset designed for detecting hallucinations in domain-specific question-answering tasks, providing evaluation metrics for retrieval-augmented LLMs.

- **DAHL** Seo et al. (2024): A comprehensive benchmark for evaluating hallucinations in biomedical text generation, featuring atomic unit decomposition and the DAHL Score metric.

2.3 Hallucinations in Multimodal Settings

2.3.1 Definition and Challenges

Multimodal hallucinations occur when models generate outputs inconsistent with visual or auditory inputs. This phenomenon is particularly critical in applications like video understanding, where temporal and spatial accuracy are essential Bai et al. (2024).

2.3.2 Evaluation Frameworks

- **VidHalluc** Li et al. (2024): A specialized benchmark for evaluating temporal hallucinations in video understanding, assessing multiple dimensions including action recognition and scene transitions.
- **MHaluBench** Chen et al. (2024): A meta-evaluation framework for comprehensive multimodal hallucination detection across diverse categories.

2.4 Hallucinations in Natural Language Generation

In natural language generation tasks such as dialogue generation, abstractive summarization, and neural machine translation, hallucinations frequently manifest as plausible but factually incorrect outputs Ji et al. (2023). These inaccuracies significantly impact the reliability and trustworthiness of these models.

2.5 Hallucinations in Academic Reference Generation

Academic reference generation represents a critical challenge, with studies demonstrating that even state-of-the-art models frequently generate fabricated or inaccurate citations Agrawal et al. (2024). This limitation underscores the urgent need for continued research in hallucination mitigation, particularly in tasks where factual accuracy is paramount. In addressing these challenges, our work specifically focuses on *Domain-Specific Biases* by leveraging over 150 categories of papers from the ArXiv repository, providing a comprehensive evaluation across diverse academic domains.

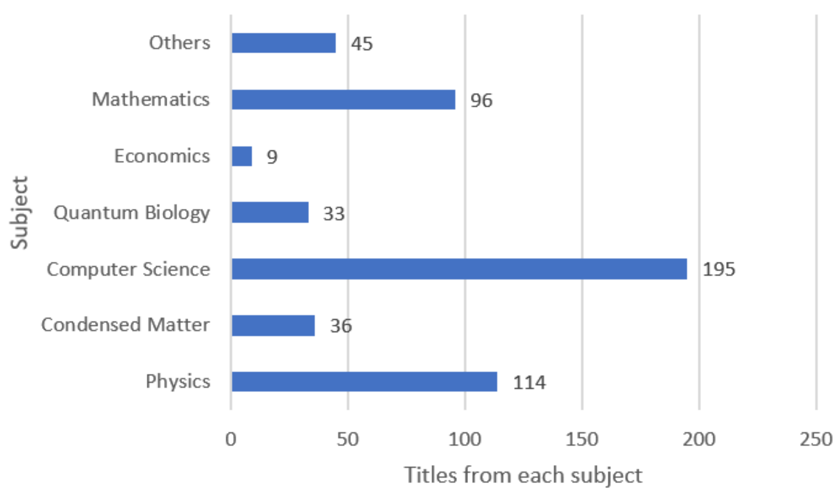

Figure 1: Number of titles from each subject.

3 Dataset

Here, we describe the datasets used for the **Jumbled Title** task and the **Mixed Title** task. A total of 176 categories were used to construct the dataset for both tasks. We used 3 titles from each category giving us 528 titles. Figure 1 shows the number of titles from each subject, Computer Science had the most number of categories at 65 having a total of 195 titles. Economics had the least at 3 categories and 9 titles.

In Table 3 we provide a further insights into the dataset for both the Jumbled Titles and Mixed Titles tasks. Using the Flesch reading ease score to Wikimedia projects (2004a), we find that Jumbled Titles falls in the `Very difficult to read. Best understood by university graduates` range, while Mixed Titles is categorized as `Extremely difficult to read. Best understood by university graduates`. Using the Gunning fog index to Wikimedia projects (2004b), we find Jumbled Titles scoring 17 falls into the `College graduate` level. Mixed Titles scores 19, which exceeds the standard Gunning fog categories but is most appropriately classified as `College graduate` level. These scores further validate the technical complexity of our dataset. The dataset exhibits diverse title lengths, with Jumbled Titles ranging from 2-24 words and Mixed Titles from 8-33 words. This broad distribution ensures robust evaluation across varying input complexities.

The **Jumbled Title** task includes 528 jumbled titles, distributed among the 176 categories of the ArXiv dataset, Table 1 presents a few examples of the dataset used for the Jumbled Title task.

Jumbled Title	**Original Title**
Hydrodynamic bubble to obstruction expansion	Hydrodynamic obstruction to bubble expansion
with Warm Microwave Background Constraining Inflation Cosmic the	Constraining Warm Inflation with the Cosmic Microwave Background
QCD Hadron Colliders Three-Jet Corrections Production Two-Loop at for Leading-Color	Leading-Color Two-Loop QCD Corrections for Three-Jet Production at Hadron Colliders
enumeration theorem polynomials Order Pólya's and	Order polynomials and Pólya's enumeration theorem

Table 1: Examples from the dataset used for the Jumbled Title task

The **Mixed Title** task consists of 265 mixed titles, derived from the 176 categories in the ArXiv dataset. Table 2 shows examples from the dataset used for the Mixed Title task.

Mixed Title	**Title 1**	**Title 2**
Bioconvection Transport Irradiation Suspensions: across Non-scattering Coarse-grain Molecular under Heating Fullerene in Membrane Collimated Above a Study Phototactic from Cell Dynamics of	Heating from Above in Non-scattering Suspensions: Phototactic Bioconvection under Collimated Irradiation	Coarse-grain Molecular Dynamics Study of Fullerene Transport across a Cell Membrane
Value of oil and gas Semi-intrusive exchange in the uncertainty quantity multiscale for stock gas london and change disclosures components oil of the relevance models propagation of upstream reserve companies	Value relevance of the components of oil and gas reserve quantity change disclosures of upstream oil and gas companies in the London Stock Exchange	Semi-intrusive uncertainty propagation for multiscale models

Table 2: Overview of the dataset for the Mixed Title task

Task	**Total**	**Categories Count**	**Avg. Len**	**Range**	**Readability**	
					Gunning Fog	**Flesch**
Jumbled Titles	528	176(3)	9.45	2-24	17	16
Mixed Titles	265	176(3)	18.89	8-33	20	8

Table 3: Dataset Statistics for Jumbled and Mixed Titles Tasks

4 Methodology

Our pipeline consists of two main tasks as seen in Figure 2. With our prompts, we aim to demonstrate and experiment with how users would typically interact with these models in real-world scenarios. This approach reflects the reality that average users, and even expert users, rarely invest time in fine-tuning their prompts with elaborate instructions during routine interactions.

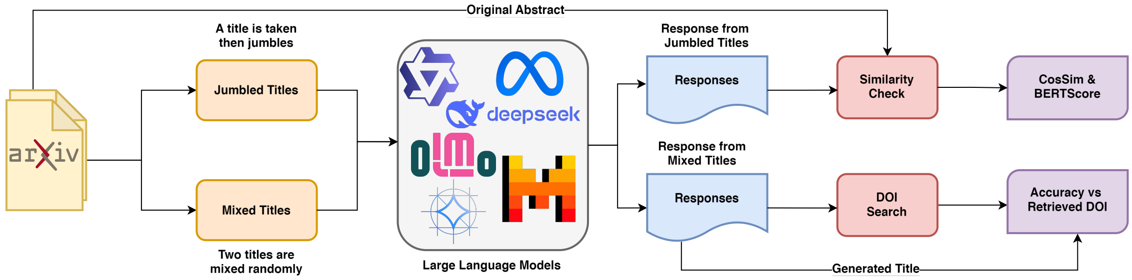

Figure 2: Our proposed pipeline for evaluating the LMs on use of the ArXiv as source for papers.

4.1 Jumbled Titles

We select 5 titles from each category available in the ArXiv dataset. Each title is scrambled within itself, creating a jumbled version of the original title. These jumbled titles serve as input to the LLMs, which are prompted using the following template:

```
Tell me about this research paper: [title]
```

Where *[title]* represents the jumbled title.

Algorithm 1 describes how the dataset is built for the Jumbled Titles task. Once the language models generate their responses, we evaluate the similarity of these responses against the original abstract of the corresponding paper. This evaluation is performed using the *all-MiniLM-L6-v2* model [Reimers & Gurevych (2019)] to generate embeddings. To quantify the degree of alignment between the generated response and the original abstract, cosine similarity scores are calculated. We also use BERTScore[Zhang et al. (2020)] and Semantic Textual Similarity (STS) Reimers & Gurevych (2019) as additional metrics for evaluating similarities.

Algorithm 1 Create Jumbled Titles Dataset

Require: Parquet file path *parquet_file_path*, Output CSV file path *output_csv_file_path*
Ensure: CSV file with jumbled titles is saved

function JUMBLE_TITLE(title)
  Split the *title* into words
  Randomly shuffle the words
  Return the shuffled words joined into a single string
end function
procedure CREATE_JUMBLED_TITLES_DATASET(parquet_file_path, output_csv_file_path)
  Load the dataset from *parquet_file_path* into DataFrame *df*
  Apply the *jumble_title* function to the *title* column in *df*
  Create a new DataFrame *jumbled_titles_df* with the jumbled titles
  Save *jumbled_titles_df* to CSV file at *output_csv_file_path* without the index
end procedure

4.2 Mixed Titles

In the second part of the pipeline, we create mixed titles by combining two random paper titles from the dataset. These mixed titles are then used as input to the LLMs with the following prompt template:

```
Tell me 2 papers related to this and only mention the Title and the DOI: [title]
```

Where *[title]* represents the mixed title.

Algorithm 2 shows how the Mixed Titles dataset is built. After the language models generate their responses, we perform a detailed evaluation of the provided DOIs using the *Crossref API*, *DataCite API*, *UnPaywall API* and *OpenAlex API* to ensure thorough search of all possible DOIs is done. The evaluation process involves the following two steps:

1. **Checking DOI Validity:** For each model-generated DOI, we verify its existence through API requests to established academic databases. This validation step is crucial for assessing the model's ability to generate legitimate academic identifiers.

2. **Verifying Title Accuracy:** For validated DOIs, we retrieve the official paper titles from the corresponding databases and compare them against the model-generated titles. This comparison enables us to evaluate the model's accuracy in associating DOIs with their correct academic works.

Algorithm 2 Create Mixed Titles Dataset

Require: Parquet file path *parquet_file_path*, Output CSV file path *output_csv_file_path*
Ensure: CSV file with mixed titles is saved

```
function MIX_TITLES(title1, title2)
    Split title1 and title2 into words
    Concatenate the words from both titles into a list mixed_words
    Randomly shuffle mixed_words
    Return the shuffled words joined into a single string
end function
procedure CREATE_MIXED_TITLES_DATASET(parquet_file_path, output_csv_file_path)
    Load the dataset from parquet_file_path into DataFrame df
    Extract the title column from df and convert it to a list titles
    Randomly shuffle the titles list
    if the length of titles is odd then
        Append an empty string to titles
    end if
    Initialize an empty list mixed_titles_data
    for each pair of titles in titles do
        Mix the pair of titles using the mix_titles function
        Append a dictionary with keys mixed_title, title1, and title2 to mixed_titles_data
    end for
    Create a new DataFrame mixed_titles_df from mixed_titles_data
end procedure
```

5 Results

To run the inference on the models we used 2× T4 Tesla [16GB]. We made use of PyTorch Paszke et al. (2019) and Huggingface's Transformers Wolf et al. (2020). It took us approximately 2.5 to 3 hrs on average to run the entire pipeline for each model. We also use 4-bit quantization using bitsandbytes bitsandbytes foundation (2024) for faster and efficient inference.

In Table 4, we present the performance of models on the Jumbled Titles task. Mistral v0.3 Jiang et al. (2023) achieved the highest similarity scores, with 0.605 on cosine similarity, 0.542 on BERTScore and 0.607 on STS. This was followed by Qwen2.5 (7B) being the second best performing model overall. On average, BERTScore reduced the similarity between the texts by 1.068% compared to cosine similarity. The worst

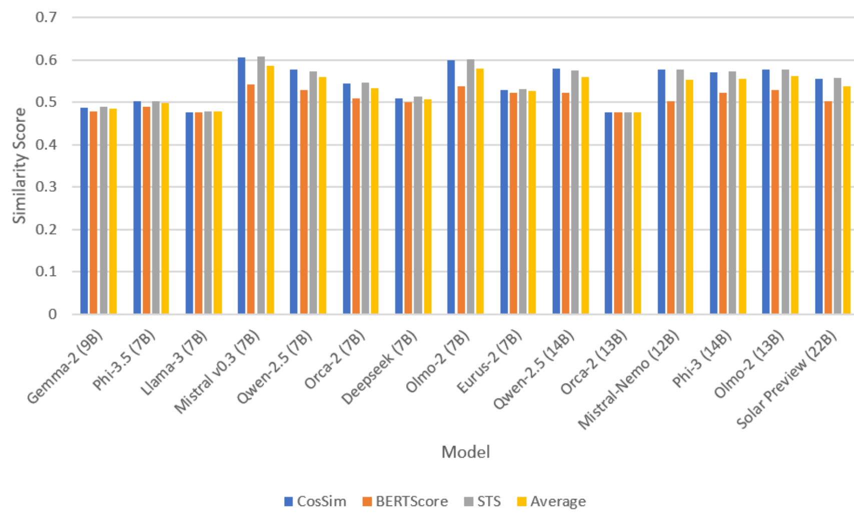

Figure 3: CosSim, BERTScore and STS Scores for all models.

Model	**Parameters**	**CosSim**	**BERTScore**	**STS**	**Average**
Gemma-2	9B	0.487	0.478	0.489	0.485
Phi-3.5	7B	0.502	0.489	0.503	0.498
Llama-3	7B	0.477	0.477	0.479	0.478
Mistral v0.3	7B	**0.605**	**0.542**	**0.607**	**0.585**
Qwen-2.5	7B	0.578	0.528	0.572	0.559
Orca-2	7B	0.544	0.508	0.546	0.533
Deepseek	7B	0.509	0.501	0.513	0.507
Olmo-2	7B	*0.600*	*0.537*	*0.602*	*0.580*
Eurus2	7B	0.528	0.523	0.530	0.527
Qwen-2.5	14B	0.579	0.522	0.575	0.559
Orca-2	13B	0.476	0.475	0.477	0.476
Mistral-Nemo	12B	0.577	0.503	0.578	0.553
Phi-3	14B	0.571	0.522	0.572	0.555
Olmo-2	13B	0.577	0.528	0.578	0.561
Solar Preview	22B	0.55	0.502	0.558	0.537

Table 4: Cosine Similarity Scores, BERTScores and STS Scores between generated response and the original abstract of the paper for various language models. Best performing model represented by **Bold** and second best by *Italics*.

performing model on all tasks was Orca-2 (13B) with cosine similarity, BERTScore and STS at 0.476, 0.475, 0.477 respectively averaging 0.476.

In Table 5, we evaluate the performance of the models on the Mixed Titles task. Among the models, Mistral-v0.3 (7B) generated the highest number of DOIs at 425, with 25.56% of the generated DOIs actually existing. Gemma-2 (9B) and Qwen-2.5 (14B) had the lowest percentage of valid DOIs, with 0% of the generated DOIs found to exist. In contrast, the model with the highest percentage of valid DOIs was the smaller Qwen-2.5 (7B), achieving a 40.70% validity rate with 70 valid DOIs generated. In Figure 5, we see a representation of the DOIs found vs DOIs not found for each model; in all cases, the number of DOIs not found is higher than that of the DOIs found. All models demonstrated a significant shortcoming: for every valid DOI retrieved, the associated title was incorrect 100% of the time when compared to the title generated by the model. It is also worth mentioning that Mistral-v0.3(7B) generated the most number of valid DOIs at 109 of the 425 generated actually existing

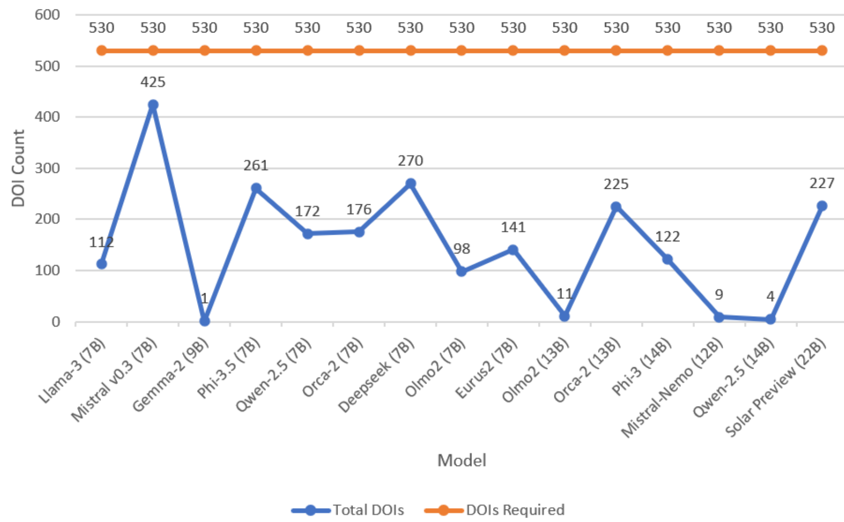

Figure 4: DOIs generated by each model when prompted during the *Mixed Title* task.

Model	Total DOIs	DOIs Found	DOIs Not Found	Matching Titles
Llama-3 (7B)	112	9 [8.04%]	8 [91.96%]	0.00%
Mistral v0.3 (7B)	425	109 [25.65%]	316 [74.35%]	0.00%
Gemma-2 (9B)	1	0 [0.00%]	1 [100.00%]	0.00%
Phi-3.5 (7B)	261	36 [13.79%]	225 [86.21%]	0.00%
Qwen-2.5 (7B)	172	70 [40.70%]	102 [59.30%]	0.00%
Orca-2 (7B)	176	20 [18.87%]	86 [81.13%]	0.00%
Deepseek (7B)	270	62 [22.96%]	208 [77.04%]	0.00%
Olmo2 (7B)	98	12 [12.24%]	86 [87.76%]	0.00%
Eurus2 (7B)	141	40 [28.37%]	101 [71.63%]	0.00%
Olmo2 (13B)	11	4 [36.36%]	7 [63.64%]	0.00%
Orca-2 (13B)	225	39 [17.33%]	186 [82.67%]	0.00%
Phi-3 (14B)	122	35 [28.69%]	87 [71.31%]	0.00%
Mistral-Nemo (12B)	9	2 [22.22%]	7 [77.78%]	0.00%
Qwen-2.5 (14B)	4	0 [0.00%]	4 [100.00%]	0.00%
Solar Preview(22B)	227	59 [25.99%]	168 [74.01%]	0.00%

Table 5: DOI Search and Title Comparison Results

6 Conclusion

This paper evaluates the extent of hallucination in state-of-the-art language models by designing two tasks: **Jumbled Titles** and **Mixed Titles**. In the Jumbled Titles task, the fifteen evaluated models achieved an average cosine similarity score of 0.544, 0.509 on BERTScore, and 0.545 on STS. Mistral-v0.3 was the best-performing model across all metrics averaging 0.585 on the Jumbled Titles task, outperforming models twice and thrice its size.

For the Mixed Titles task, while models generated DOIs for the mixed titles, they often cited non-existent papers or mismatched DOIs. These results underscore critical limitations in maintaining factual accuracy in domain-specific contexts. On average, valid DOIs were generated only 17.75% of the time. Every model also completely failed to generate the corresponding DOI for the title they generated, showing that models struggle with maintaining factual consistency. To further highlight *Prompt Adherence*, it is worth noting that none of the models generated the required number of two DOIs for each *Mixed Title*. This results in a

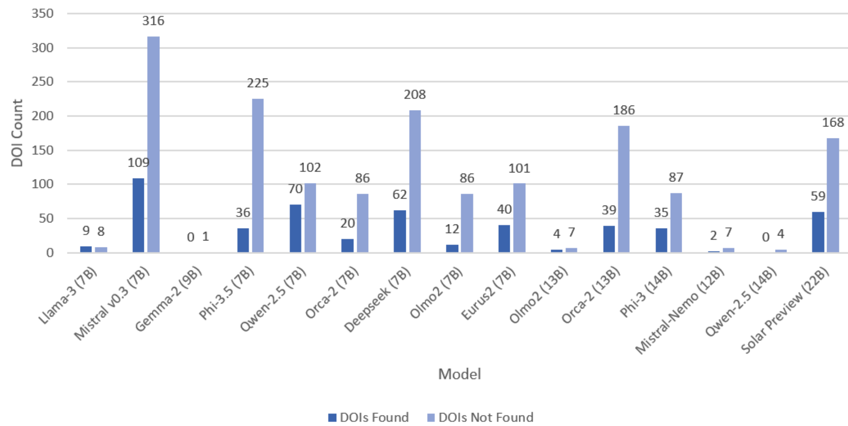

Figure 5: DOIs Found for each model vs DOIs not for each model

discrepancy, as the expected output for each model was 530 DOIs (given 265 mixed titles), but none of the models met this requirement, as seen in Figure 4.

6.1 Model Size Performance

In Table 4 and Table 5, we observe a concerning trend where many of the larger models are significantly outperformed by their smaller counterparts in both tasks. For instance, the 7B Mistral-v0.3 outperforms models up to three times its size, while the Solar Preview(22B) demonstrates mediocre performance despite its substantially larger parameter count, even for the same models with Qwen2.5 where the 7B parameter model outperforms the 14B parameter variant. This pattern raises serious questions about the relationship between model size and task performance. The results starkly highlight that simply scaling up model parameters does not guarantee better performance in specialized tasks, particularly those requiring precise adherence to instructions and factual accuracy. This counterintuitive finding challenges the common assumption that larger language models inherently perform better, suggesting that architectural choices and training approaches might be more crucial than raw parameter count for achieving superior performance.

7 Limitations

There are several limitations to our work:

1. **Model Selection**: Our evaluation focuses on smaller models due to computational constraints. Results may differ significantly with larger variants, which could exhibit different performance characteristics. Although our findings in Section 6.1 shows that might not always be the case, as we observed smaller models often outperforming their larger counterparts.

2. **Model Quantization**: We use 4-bit quantization for inference. While this may reduce performance, studies suggest the impact is minimal Jin et al. (2024). The trade-off between computational efficiency and potential performance impact was deemed acceptable for our experimental setup.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL `https://arxiv.org/abs/2404.14219`.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. Do language models know when they're hallucinating references? In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 912–928, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-eacl.62`.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2024. URL `https://arxiv.org/abs/2404.18930`.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*, 2024.

Elijah Berberette, Jack Hutchins, and Amir Sadovnik. Redefining "hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation, 2024. URL `https://arxiv.org/abs/2402.01769`.

bitsandbytes foundation. bitsandbytes-foundation/bitsandbytes: Accessible large language models via k-bit quantization for pytorch., Decemeber 2024. URL `https://github.com/bitsandbytes-foundation/bitsandbytes`. [Online; accessed 2024-12-16].

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3235–3252, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.178. URL `https://aclanthology.org/2024.acl-long.178/`.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of arxiv as a dataset, 2019. URL `https://arxiv.org/abs/1905.00075`.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL `https://arxiv.org/abs/2401.02954`.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387. URL https://aclanthology.org/2022.naacl-main.387.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey, 2023. URL https://arxiv.org/abs/2310.19736.

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. *ACM Transactions on Information Systems*, November 2024. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

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 Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On large language models' hallucination with regard to known facts. *arXiv preprint arXiv:2403.20009*, 2024.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. A comprehensive evaluation of quantization strategies for large language models, 2024. URL https://arxiv.org/abs/2402.16775.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding, 2024. URL https://arxiv.org/abs/2412.03735.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024. URL https://arxiv.org/abs/2402.06196.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. Orca 2: Teaching small language models how to reason, 2023. URL https://arxiv.org/abs/2311.11045.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. URL https://arxiv.org/abs/2307.06435.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2024. URL https://arxiv.org/abs/2501.00656.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2541–2573, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.155. URL https://aclanthology.org/2023.emnlp-main.155.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, and Zhe Feng. Delucionqa: Detecting hallucinations in domain-specific question answering, 2023. URL https://arxiv.org/abs/2312.05200.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models, 2024. URL https://arxiv.org/abs/2405.09589.

Jean Seo, Jongwon Lim, Dongjun Jang, and Hyopil Shin. Dahl: Domain-specific automated hallucination evaluation of long-form text through a benchmark dataset in biomedicine, 2024. URL https://arxiv.org/abs/2411.09255.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So. Confabulation: The surprising value of large language model hallucinations. *arXiv preprint arXiv:2406.04175*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Mistral Team. mistralai/mistral-nemo-instruct-2407 · hugging face. URL https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407. [Online; accessed 2025-01-15].

Mistral AI Team. mistralai/mistral-7b-instruct-v0.3 · hugging face, December 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. [Online; accessed 2024-12-16].

Contributors to Wikimedia projects. Flesch–kincaid readability tests - wikipedia, 7 2004a. URL https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests. [Online; accessed 2025-01-17].

Contributors to Wikimedia projects. Gunning fog index - wikipedia, 10 2004b. URL https://en.wikipedia.org/wiki/Gunning_fog_index. [Online; accessed 2025-01-17].

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024. URL https://arxiv.org/abs/2401.01313.

upstage. upstage/solar-pro-preview-instruct · hugging face, 11 2024. URL `https://huggingface.co/upstage/solar-pro-preview-instruct`. [Online; accessed 2025-01-15].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL `https://arxiv.org/abs/1910.03771`.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL `https://arxiv.org/abs/1904.09675`.