FActBench: A Benchmark for Fine-grained Automatic Evaluation of LLM-Generated Text in the Medical Domain

Anum Afzal
Technical University of Munich
anum.afzal@tum.de

Juraj Vladika
Technical University of Munich
juraj.vladika@tum.de

Florian Matthes
Technical University of Munich
matthes@tum.de

Abstract

Large Language Models tend to struggle when dealing with specialized domains. While all aspects of evaluation hold importance, factuality is the most critical one. Similarly, reliable fact-checking tools and data sources are essential for hallucination mitigation. We address these issues by providing a comprehensive Fact-checking Benchmark FActBench covering four generation tasks and six state-of-the-art Large Language Models (LLMs) for the Medical domain. We use two state-of-the-art Fact-checking techniques: Chain-of-Thought (CoT) Prompting and Natural Language Inference (NLI). Our experiments show that the fact-checking scores acquired through the Unanimous Voting of both techniques correlate best with Domain Expert Evaluation.

1 Introduction

In the quickly evolving era of Natural Language Processing (NLP), Large Language Models (LLMs) are making their way into almost all use cases and domains. In most tasks, they have shown tremendous generative capabilities and a good understanding of text. However, they still tend to hallucinate in critical domains like the Medical domain. Contemporary LLMs are typically evaluated against general benchmarks and their assessment of the Medical domain is usually lacking. While it is essential to mitigate hallucinations, as a first step some reliable automatic fact-checking indicators are needed (Clusmann et al., 2023). The field of automatic fact-checking in LLMs is rapidly developing making it essential to find trustworthy techniques and data sources.

The state-of-the-art techniques for Automatic Fact Checking include Natural Language Inference (NLI) (Mor-Lan and Levi, 2024; Akhtar et al., 2024) using DeBERTa (He et al., 2021), or through Chain-of-thought (CoT) (Wei et al., 2022) by using an LLM as a judge (Zheng et al., 2023). Given the importance of Factual correctness in a critical domain such as medicine, it is helpful to rely on more than one technique for Fact-checking. Therefore, we explore the idea of Unanimous Voting such that an atomic fact is only considered to be factually correct if it is supported by both techniques.

Hallucinations can generally be divided into input-conflicting, context-conflicting, and fact-conflicting (Zhang et al., 2023). The focus of our work lies in fact-conflicting, which is hallucination, where facts in output contradict the world knowledge. Additionally, our work builds on top of FActScore, a CoT-based approach for fact-checking. We adapt it to support user-provided grounding documents, making it suitable for tasks like RAG and Summarization. We present an automatic Fact-Checking Benchmark **FActBench**[1] with the following contributions:

- We fact-check six contemporary LLMs using Atomic Facts (Min et al., 2023) on four generations tasks: Text Summarization, Lay Summarization, Retrieval Augmented Generation (RAG), and Open-ended Generation.
- We compare Intrinsic (Grounding Document) and Extrinsic (Wikipedia Dump) Fact-checking techniques in our experiments.
- We evaluate NLI, CoT as well as Unanimous Voting (UnVot) for the final prediction using domain expert evaluations as reference.

Details about all the datasets we use can be found in their original papers, including appropriate licenses and terms of use.

2 Related Work

Hallucinations are a common problem in Natural Language Generation (NLG) tasks such as abstrac-

[1]Code for FActBench can be found at github.com/jvladika/FactSumm/

tive text summarization, generative question answering, or dialogue generation (Ji et al., 2023). Detecting hallucinations is tied to the problem of measuring the factuality of model output (Augenstein et al., 2023; Zhao et al., 2024). Hallucinations can be detected with approaches looking at the uncertainty in models' logits (Varshney et al., 2023) or with approaches that fact-check model output over external knowledge sources (Chern et al., 2023).

Some recent works approached evaluation with question answering (Scialom et al., 2021) or NLI (Utama et al., 2022). Most recent methods leverage LLMs by querying them with prompts that directly ask for a score, like G-Eval (Liu et al., 2023a), or evaluate the generated text with the veracity of its atomic facts, like FActScore (Min et al., 2023). Fadeeva et al. (2024) develop a method that does not require external knowledge for fact-checking as they leverage token-level uncertainty to identify the potentially factually incorrect generated section in the output. Similarly, Sankararaman et al. (2024) introduces Provenance, a technique that uses NLI models to check if the RAG output is factually correct with reference to context. Lastly, Chen et al. (2024) present FactCHD, a benchmarking for fact-conflicting hallucination detection for the General, Scientific, Health, and COVID-19 domains.

3 FActBench: Benchmark

In our Benchmark, we use two SotA techniques, NLI and CoT, to evaluate 6 models on 4 different tasks. We follow the approach introduced by Min et al. (2023) to break all generations into a list of atomic facts which are then used for fact-checking. Since all our tasks with the exception of Open Generation, use a source document for grounding, we opt for a hybrid approach such that we first perform fact-checking using an intrinsic approach, followed by an extrinsic one.[2] The latter only performs evaluation on atomic facts that have been marked as hallucinations in the first step. We employ such an approach because it is possible for an atomic fact to be factually correct as per the world knowledge, even if it is not supported by the grounding document. We show this methodology in Figure 1 that illustrates how different fact-checking techniques and data sources interact with each other.

[2]In factuality evaluation, *intrinsic* hallucinations are those that contradict the reference document, while *extrinsic* hallucinations are those that contradict the external world knowledge.

3.1 Techniques

Baseline: FActScore As a baseline, we first report on task performance using the established FActScore metric, following their external checks on Wikipedia with no grounding document. The reason we use it is its popularity in papers involving generative NLP tasks in the last couple of years (Dhuliawala et al., 2024; Chang et al., 2024; Huang et al., 2025). Later, our goal is to show that the combination of methods we use instead of raw FActScore lead to a more faithful evaluation framework and a better alignment with human scores.

Natural Language Inference (NLI): We utilize NLI as the first evaluation method. NLI aims to predict the logical relation between a premise and a hypothesis, including entailment, contradiction, and a neutral stance. We use the generated answer as the premise and the reference answer as the hypothesis. The intuition behind this approach is that a good answer should logically entail the reference. NLI has been applied for evaluating the quality of summaries and text generation (Mishra et al., 2021; Laban et al., 2022; Steen et al., 2023).

Following this approach, we use DeBERTa-v3 (He et al., 2023), shown to work well with NLI and reasoning tasks. We use the version *Tasksource*, fine-tuned on a wide array of NLI & classification datasets, which works well with long inputs (Sileo, 2023).[3] We take *entailment* predictions as a sign of the atomic fact being supported by the original text and *contradiction* as a sign of hallucination. We additionally check the contradicting atomic facts in an extrinsic way, by predicting their NLI class with the relevant Wikipedia context as the hypothesis.

Chain-of Thought (CoT) Prompting: For evaluation using Chain-of-Thought Prompting, we adapted FActScore, an existing CoT-based fact-checking tool. This technique is suitable for open-ended generation and uses a Wikipedia dump as the knowledge source. FActScore supports extrinsic fact-checking by retrieving the most relevant passages from Wikipedia using user-defined topics. We adapt FActScore to support external documents as the basis for fact-checking. This "topic" should be the name of a real Wikipedia article, from which the relevant passages are retrieved. We also include a LLM-based topic generator so it is not required to manually define the topic when evaluating using

[3]https://huggingface.co/tasksource/deberta-base-long-nli

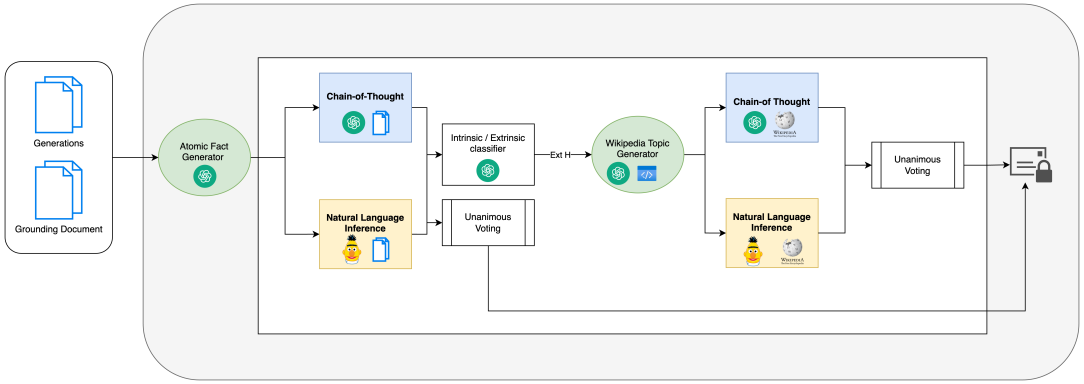

Figure 1: Block Diagram depicting how different fact-checking techniques interact with different data sources. Chain-of-Thought uses an LLM whereas Natural Language Inference uses a small LM as the backbone.

passages from the Wikipedia dump. We use GPT-4o mini as the backbone of FActScore+, which serves as a compromise between cost and quality.

Unanimous Voting (UnVot): To produce a reliable fact-checking approach, we explore the idea of Unanimous Voting. This means we only consider an atomic fact to be correct if both NLI and CoT support it. This technique is especially useful for applications where high precision is needed.

Human Evaluation: We evaluate CoT, NLI, and UnVot techniques by correlating to domain expert judgment. We recruited 8 in-house employed individuals with a medical background to serve as annotators. A random subset of 80 generations (20 per task) was manually annotated such that each generation was evaluated by two annotators. They were instructed to follow the same hybrid, using both the original article and Wikipedia as a basis for fact-checking. Annotators were asked to assign a score between 1 and 100 to the generation estimating the factual correctness of the text.

3.2 Tasks

We include four tasks in our Benchmark, including Text Summarization, Lay Summarization, Retrieval Augmented Generation, and Open-ended Generation. The prompts used for all four tasks are shown in Appendix A. We summarize the datasets used for the tasks in Table 1 and discuss them below. All the datasets can be found in their respective original papers, together with appropriate licenses.

Text Summarization. This task refers to the ability of an LLM to summarize a long scientific article into a summary. We used 1000 random samples from the PubMed Summarization dataset (Cohan et al., 2018), which is derived from the original PubMed dump.

Lay Summarization. Contrary to normal text summarization, Lay Summarization refers to the model's ability to create a layman summary of biomedical articles. We use 1000 random samples from the PLOS dataset introduced by Goldsack et al. (2022).

Retrieval Augmented Generation (RAG). We use BioASQ-QA (Krithara et al., 2023), a biomedical question answering (QA) dataset designed to reflect the real information needs of biomedical experts. The questions are written by experts and evidence comes from PubMed. We use the *summary* subset – 1130 questions paired with human-selected evidence snippets from PubMed and human-written "ideal answers" based on those snippets. We use the gold snippets as input to an LLM and prompt it to generate an answer to the given question, thus simulating a **RAG** pipeline.

Open-ended Generation. In this setting, no context is used and the model is prompted to generate an answer based on its knowledge. We again use the BioASQ dataset from the RAG task – we take the 1130 questions and use them as input to an LLM by prompting it to answer the question.

Task	Dataset	#Source W	#Gen W
Summ	PubMed	3,053.9	256
Lay Summ	PLOS	6,696.8	256
RAG	BioASQ-QA	351.9	116.5
Gen	BioASQ-QA	351.9	default

Table 1: Average word count of articles (#W) and # generation tokens (#Gen W) during inference for tasks with respective datasets. Summ = Text Summarization, Lay Summ = Layman Summarization, RAG = Retrieval Augmented Generation, Gen = Open-ended Generation.

Models	Summarization			Lay Summarization			RAG (QA)			Open-ended Gen		
	CoT	NLI	UnVot	CoT	NLI	UnVot	CoT	NLI	UnVot	CoT	NLI	UnVot
FActBench (Grounding Document)												
GPT-4o mini	95.8	77.4	86.6	95.4	94.8	95.1	25.4	**77.7**	51.5	44.5	**50.4**	47.5
Llama3.1 8b	95.3	87.8	85.28	95.4	93.5	94.4	35.6	76.7	56.1	74.4	35.6	55.0
Llama3.1 70b	**96.52**	84.59	82.84	96.1	94.1	95.1	31.3	76.3	53.8	37.3	46.5	41.8
Mistral 7b	95.8	82.55	80.38	96.3	97.32	94.75	82.9	73.1	78.0	80.9	32.5	56.6
Mixtral 8 x 7b	95.2	87.86	**95.5**	96.5	97.0	95.0	**88.2**	75.0	**81.6**	**85.4**	36.9	**61.1**
Gemma 9b	84.55	71.95	68.77	82.94	80.65	75.48	35.8	44.0	43.7	54.1	30.5	28.0
FActBench (Grounding Document + Wikipedia)												
GPT-4o mini	96.8	82.6	80.4	96.2	96.6	93.4	97.3	**78.2**	**76.4**	**95.8**	**51.4**	**50.3**
Llama3.1 8b	96.4	88.85	86.25	96.5	94.2	91.5	98.2	77.1	76.1	79.3	36.7	32.1
Llama3.1 70b	97.27	85.71	83.9	97.0	94.8	92.0	97.2	76.8	75.1	90.9	47.7	45.9
Mistral 7b	96.51	83.59	81.34	**97.83**	**96.7**	**94.93**	**98.6**	73.5	72.7	92.1	33.2	31.9
Mixtral 8 x 7b	96.9	88.68	**86.24**	97.5	97.2	95.1	97.7	75.3	74.0	93.0	37.8	36.5
Gemma 9b	93.03	74.46	70.99	91.11	81.68	76.43	97.4	45.0	44.6	80.1	31.5	28.8
Baseline: FActScore (Wikipedia)												
GPT-4o mini		51.34			52.6			19.4			41.4	
Llama3.1 8b		43.97			49.4			25.3			71.3	
Llama3.1 70b		50.08			48.8			24.0			34.8	
Mistral 7b		46.11			50.02			61.1			78.4	
Mixtral 8 x 7b		49.71			51.00			**64.5**			**81.6**	
Gemma 9b		**53.54**			**54.56**			44.0			52.0	

Table 2: Factchecking scores of six LLMs on four tasks using Chain-of-Thought (CoT) prompting, Natural Language Inference (NLI), and Unanimous voting (UnVot). We show scores by incorporating two different knowledge sources.

3.3 Models

We include six LLMs in our experiments including `Llama3.1 8b` (Dubey et al., 2024) `Llama3.1 70b`, `Mistral 7b` (Jiang et al., 2023), `Mixtral 8x7b` (Jiang et al., 2024a), `Gemma2 9b` (Team et al., 2024) and lastly, closed-source `GPT-4o mini`. We provide the checkpoints and technical details in Appendix B.

4 Results & Discussion

4.1 Correlation with Human Evaluation

Before discussing the benchmark results, we check the effectiveness of the techniques used. We performed human evaluation using the process described in subsection 3.1. The average fact-checking scores using the baseline, 3 techniques, as well domain expert annotations are in Table 3. The final Cohen's inter-annotator agreement κ is 0.75, which signifies substantial agreement. The baseline technique (FActScore) that uses only Wikipedia as the knowledge source severely underestimates the correctness of the generated text whereas the Chain-of-Thought technique that uses Grounding Document and Wikipedia overestimates it. Overall, it can be seen that our UnVot score derived through joint decisions of CoT and NLI correlates best with domain expert judgment. Still, it is important to point out that this holds true for the summarization, lay summarization, and RAG tasks, while the pure generation task best correlated with baseline FActScore system.

The high correlation of UnVot with human judgment is an important finding. Hiring human annotators, especially domain experts, can be a very expensive and time-consuming process. Having a metric that highly correlates with human scoring intuition can provide a good enough substitute for situations where finding human annotators is infeasible or impossible for certain labs, groups, and application use cases. A lot of focus of recent LLM research is put on aligning LLMs with human values and intuition (Wang et al., 2023), and recent LLM-as-judge evaluation metrics like G-Eval (Liu et al., 2023b), Prometheus (Kim et al., 2024), and TIGERScore (Jiang et al., 2024b) put a high emphasis on the correlation of their metrics with humans. As future work, it would be interesting to compare these metrics with UnVot as well, which we currently skip due to resource constraints.

4.2 Task and LLM Performance

We summarize the Fact-checking scores in Table 2, which show that the grounding helps LLMs to be more truthful. In terms of tasks, LLMs tend to hallucinate more when prompted to do open-ended

Task	Baseline	CoT*	NLI*	UnVot*	Human
Summ	54.81	96.87	85.41	83.45	84.0
LaySumm	52.5	97.6	91.09	88.94	88.7
RAG	38.43	100.0	83.04	83.04	87.3
PureGen	71.26	88.17	31.61	31.31	62.7

Table 3: Fact-checking scores on FActScore (Baseline), Chain-of-Thought (CoT), Natural Language Inference (NLI), Unanimous Voting (UnVot), and Domain Expert Evaluation (Human). * refers to final scores with intrinsic followed by extrinsic fact-checking.

generation in the medical domain. However, the performance on other grounding-based task show that given the correct context and supporting document, LLMs are good at understanding a complex domain such as the medical domain. Within each task, LLM performance is mostly uniform. As expected, Open-ended generation is the most challenging task, which is expected due to the LLM using its internal knowledge to answer questions, which can lead to hallucinations. Lay summarization was the most factually correct task, likely owing to the nature of lay text where simpler terms and phrasing is used, which reduces the possibility of mixing up complex scientific terms with one another, which would lead to hallucinations.

Surprisingly, we see no big difference in models with respect to their sizes. However, both `Mistral` and `Mixtral` lead the board for two summarization tasks. While `Mixtral` performs best for two QA tasks with only the grounding document, `GPT` comes on top after extrinsic checks, showing its high awareness of Wikipedia in pre-trained knowledge. Two `Llama` models come close to `Mixtral`, while `Gemma` performs the worst on all tasks.

5 Conclusion and Future Work

We present a Benchmark providing insights over contemporary LLMs across 4 tasks in the medical domain. We discuss Chain-of-Thought, Natural Language Inference, and Unanimous Voting as fact-checking techniques. Through Domain Expert Evaluation, we show the Unanimous Voting technique to be most reliable. We also explored the effectiveness of two knowledge sources, namely a Grounding Document and Wikipedia, for evaluation and found that using more than one knowledge source leads to an increase in factuality scores. Lastly, we found that LLMs are mostly factually incorrect for Open-ended generation in the medical domain and tend to be more faithful for tasks like Summarization and RAG, where some context is provided to the LLM for generation. We envision our evaluation benchmark to be easily applied for fact-checking across other domains in future.

Limitation

Due to the high computation costs, we use only one model as the backbone for each factuality evaluation technique. Even though we evaluated six Large Language Models on four diverse tasks, these tasks may not be enough to capture the entirety of LLM performance and the quickly evolving landscape of new models.

Additionally, our two evaluation techniques with NLI and FactScore+ CoT are not perfect and it is possible there were incorrect predictions of which facts were supported or refuted by evidence. Even though our manual inspection and human evaluation showed a good correlation with automated metrics, there will always be some mishaps and incorrect verdicts.

Finally, our approach relies on making numerous calls to the external API and to the Wikipedia dump database instance in case of extrinsic fact-checking, which can all slow down the overall pipeline. An alternative would have been running locally hosted open-source models, but this was out of our budget due to computational costs. Future work could explore these solutions and make the process faster.

Ethics Statement

Throughout our experiments, we strictly adhere to the ACL Code of Ethics. The manual evaluation was performed by in-house annotators who received a full salary, and their annotation were stored anonymously, mitigating any privacy concerns. They were informed about the task and usability of data in the research. The goal of the research is to evaluate existing techniques and introduce a new technique that can be used for fact-checking LLM generated text on four tasks in the medical domain. We use the LLMs through inference using open-source dataset and do not include in any information in model weights. The discussions and results in this paper are meant to further promote research in the area of LLM Fact-checking as well as create more awareness about their applications in the medical domain. All scripts will be made available to the research community.

References

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev2r: Evaluating evidence retrieval in automated fact-checking. *Preprint*, arXiv:2411.05375.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. *Preprint*, arXiv:2310.05189.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2024. Factchd: Benchmarking fact-conflicting hallucination detection. *Preprint*, arXiv:2310.12086.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, 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, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, 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, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, 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, 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, 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, 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, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, 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, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, 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, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, 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, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, 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, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, 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, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, 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, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, 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, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, 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, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, 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, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024b. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1).

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

Guy Mor-Lan and Effi Levi. 2024. Exploring factual entailment with NLI: A news media study. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 190–199, Mexico City, Mexico. Association for Computational Linguistics.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. 2024. Provenance: A light-weight fact-checker for retrieval augmented LLM generation output. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1305–1313, Miami, Florida, US. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, Patrick Gallinari, et al. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604. Association for Computational Linguistics.

Damien Sileo. 2023. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, NLI models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924, Toronto, Canada. Association for Computational Linguistics.

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 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. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Prasetya Utama, Joshua Bambrick, Nafise Sadat Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level nli examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *Preprint*, arXiv:2307.03987.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao 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.

A LLM Prompts:

The prompts used for LLM inferences on all four tasks are illustrated in Table 4.

TEXT SUMMARIZATION PROMPT
Summarize the given article by including the following key points: Objective: What is the main research question or objective of the study? Background: What is the context or rationale for the study? Methods: What study design, population, and methodologies were used? Key Findings: What are the most significant results or discoveries from the study? Conclusions: What conclusions do the authors draw from their findings? Clinical Relevance: How might the study's findings impact medical practice or patient care? Scientific Article: article Summary:
LAY SUMMARIZATION PROMPT
You will be provided a scientific article. Your task is to write a lay summary that accurately conveys the background, methods, key findings, and significance of the research in non-technical language understandable to a general audience. Guidelines for crafting a lay summary: Craft a detailed summary that explains the research findings and their implications, providing thorough explanations where necessary. Ensure factual accuracy and alignment with the research presented in the abstract, elaborating on key points and methodologies. Highlight the main findings and their implications for real-world scenarios, delving into specific mechanisms or methodologies used in the study and their broader significance. Incorporate descriptive language to explain complex concepts. Maintain a balanced tone that is informative and engaging, avoiding technical jargon or overly formal language. Ensure the summary provides sufficient depth and context to guide the reader through the research journey and address potential questions or areas of confusion. Scientific Article: article Summary:
RETRIEVAL AUGMENTED GENERATION PROMPT
Give a simple answer to the question based on the provided context. QUESTION: question CONTEXT: context
OPEN-ENDED GENERATION PROMPT
Give a simple answer to the question based on your best knowledge. QUESTION: question

Table 4: The prompt in the Benchmark for LLM generation output for all tasks.

B Technical Details

B.1 LLM Generations

The inference procedure was done Together AI Inference[4]. We used the instruct-tuned or chat versions of the models. As for GPT-4o mini, we used the OpenAI API and the latest snapshot available, `gpt-4o-mini` from Sep 13th, 2024. The checkpoints used for LLM inferences of the open-source models using Together AI are summarized in Table 5.

Model	**checkpoint**
`Llama 3.1 8b`	meta-llama/Meta-Llama-3.1-8B-Instruct-Turbo
`Llama 3.1 70b`	meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo
`Mistral 7b`	mistralai/Mistral-7B-Instruct-v0.3
`Mixtral 8x7b`	mistralai/Mixtral-8x7B-Instruct-v0.1
`Gemma 2 9b`	google/gemma-2-9b-it

Table 5: Together AI checkpoints of all LLMs that were used during Inferences.

B.2 Benchmark Computations

We used the OpenAI API[5] and the latest snapshot available, `GPT-4o-mini` from Sep 13th, 2024 for Fact-checking using Chain-of-Thought prompting. We leveraged Nvidia V100-16GB and Nvidia A100-80GB GPUs for performing fact-checking.

[4] https://www.together.ai/

[5] https://platform.openai.com/