Can Large Language Models Accurately Generate Answer Keys for Health-related Questions?

Davis Bartles, Deepak Gupta, and Dina Demner-Fushman
National Library of Medicine, NIH
firstname.lastname@nih.gov

Abstract

The evaluation of text generated by large language models (LLMs) remains a challenge for question answering, retrieval augmented generation (RAG), summarization, and many other natural language processing tasks. Evaluating the factuality of LLM-generated responses is particularly important in medical question answering, where the stakes are high. One method of evaluating the factuality of text is through the use of information nuggets (answer keys). Nuggets are text representing atomic facts that may be used by an assessor to make a binary decision as to whether the fact represented by said nugget is contained in an answer. Given that manual nugget extraction is expensive and time-consuming, recent RAG shared task evaluations have explored automating the nuggetization of text with LLMs. In this work, we explore several approaches to nugget generation for medical question answering and evaluate their alignment with expert human nugget generation. We find providing an example and extracting nuggets from an answer to be the best approach to nuggetization. While, overall, we found the capabilities of LLMs to distill atomic facts limited, Llama 3.3 performed the best out of the models we tested.

1 Introduction

Evaluation of automatically generated answers is a major bottleneck in the development of question-answering approaches. Although a need for new evaluation approaches was noticed as soon as the system-generated answers became more complex and abstract (Chen et al., 2019), to date, there are no widely accepted evaluation metrics to approximate human judgments on the quality and other aspects of the generated answers. The TREC 2024 Retrieval Augmented Generation track evaluation (Pradeep et al., 2024) revisited the nugget-based evaluation originally developed for judging answers to definition questions in the 2003 question answering track (Voorhees, 2004). Briefly, a nugget-based evaluation is a two-step process. In the first step, the assessors create a list of atomic facts (nuggets) that must be present in an answer for the answer to be judged correct and complete. In the second step, the assessors manually map each statement in a system-generated answer to the nuggets. Various performance metrics may then be computed. For example, the TREC 2024 RAG track labeled nuggets as *supported*, *partially supported*, or *not supported* by the answer and then computed the system scores by summing the scores of all nuggets and dividing the sum by the number of nuggets. Most importantly, the evaluation has shown correlations between scores derived from an automatic nugget evaluation and a manual nugget evaluation. This supports the belief that LLMs can be used to support evaluations as long as LLMs are not generating the ground truth (Soboroff, 2024).

To that end, we explore various approaches to the first step of the evaluation - LLMs' ability to generate atomic factual statements (*cf.* Table 1). This is particularly important in medical question answering, where, although infrequently, as demonstrated by the results of the 2024 TREC BioGen track evaluation, generated answers may contain inappropriate and potentially harmful information (Gupta et al., 2024).

The contributions of this work are as follows:

- We manually generate nuggets for the 2024 BioGen track topics.
- We propose a series of automated nugget-generation approaches considering the question, answer, and relevant documents.
- We evaluated the capabilities of LLMs' to generate nuggets.

Related work: Nugget generation could be viewed as a form of outline generation in two different settings: 1) a model or a person generating the nuggets has access to a set of answers or docu-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 354–368
July 27 - August 1, 2025

Query: *"What will mutation in runx2 affect in the future?"*	
Answer: *"The effect of the runx2 mutation depends on the kind of the mutation. This gene mostly affects bone development. Mutations can cause bone deformities, height lower than expected, extra teeth and other dental problems..."*	
Manually Extracted Nuggets:	
1. Affects (runx2 mutation, bone development)	
2. Cause (runx2 mutation, bone deformities)	
3. Cause (runx2 mutation, height lower than expected)	
4. Cause (runx2 mutation, extra teeth)	
5. Cause (runx2 mutation, dental problems) ...	
LLM (Llama 3.3) Generated Nuggets:	
1. Runx2 mutation affects bone development	
2. Runx2 mutation can cause bone deformities	
3. Runx2 mutation can result in lower than ...	
4. Runx2 mutation can lead to extra teeth	
5. Runx2 mutation can cause other dental problems	

Table 1: An example of nuggets extracted by a human and LLM for the same query and answer pair.

ments containing information needed to generate the answer; 2) the model or person are provided only with the question and have to generate the outline using their background knowledge. While, to the best of our knowledge, work on direct nugget generation is limited to the above RAG evaluation and an evaluation in which the initial set of test nuggets is generated using ChatGPT (Dietz, 2024; Farzi and Dietz, 2024), the related work on outline generation includes story generation (Wang and Kreminski, 2024) and natural language outline for code generation (Shi et al., 2024). For medical question answering, nugget-based evaluation was revisited in the evaluation of answers to questions about COVID-19 asked by patients and clinicians (Goodwin et al., 2022). The nuggets were generated manually in this evaluation. Other evaluations that leverage fact extraction were proposed for questions about biographies (Min et al., 2023) and medical question answering (Wang et al., 2024).

2 Methods

2.1 Manual Nugget Generation

For the purpose of having ground truths to evaluate LLM generated nuggets against, we provide expert-curated, manually generated nuggets for the 2024 BioGen track topics. Nuggets were captured from 40 ground truth answers. Each nugget was captured as a semantic triplet in Predicate (subject, object) form. Nuggets were identified by manually assessing the atomic facts represented in each sentence and their corresponding predicate, subject, and object. Some sentences may contain multiple atomic facts, for example, sentences comprising multiple phrases conjoined by a coordinating conjunction or lists. In such cases, the semantic triple is identified for each phrase or item in the list and recorded separately. Predicates were normalized across the dataset by mapping to a list of expert-curated predicates deemed to be complete in their coverage of the dataset and in conveying represented facts. Each medical concept contained in either the subject or object was associated with a Concept Unique Identifier (CUI) from the Unified Medical Language System (UMLS) (Lindberg et al., 1993). These associations were made by manually assessing the closest match, if any, from the UMLS Metathesaurus Browser. Some facts required more complex nugget structures including, but not limited to, "if, then" clauses and comparisons. These nuggets preserve the underlying logical structure from the answer. We generated a total of 498 nuggets from 40 question-answer pairs which has an average of 12.45 nuggets. Each nugget was reviewed by at least two reviewers.

2.2 LLM-based Nugget Generation

Model Architectures: We tested both popular open-source and proprietary models for nugget generation. The list of models includes Llama (Grattafiori et al., 2024), Gemma (Riviere et al., 2024), Mistral (Jiang et al., 2023), Phi (Abdin et al., 2024), Qwen (Qwen et al., 2025), Vicuna (Chiang et al., 2023), Falcon (Almazrouei et al., 2023), DeepSeek (DeepSeek-AI et al., 2025), GPT (OpenAI et al., 2024), Gemini (Team et al., 2023), and Claude[1]. For some families of models, we included both larger and smaller versions.

Generation Strategies: We developed extensive strategies to generate the nuggets by considering different inputs to the LLMs. Specifically, we used questions, reference answers, and cited documents provided for each assertion in the reference answer. We instructed the models to generate the appropriate nuggets. The detailed strategies are as follows: **(1) Question:** In the first strategy, we only pro-

[1] https://www.anthropic.com/claude/sonnet

vide the question to the LLMs and instruct them to generate all pertinent nuggets that directly address the user's query. We started with the zero-shot approach and extended our experiments to the few-shot approach to enable in-context learning, where we provide an example question and corresponding nuggets in the prompt to direct the model toward better performance. We call this strategy Q_0 (zero-shot) and Q_1 (one-shot).

(2) Question + Answer: We aim to assess LLMs' capability of distilling nuggets from the ground-truth answers. We hypothesized that LLMs are expected to perform well in this setting and it can be considered an upper bound for the first strategy. Similar to the first strategy, we devise two strategies (zero and one-shot) and call them QA_0 and QA_1.

(3) Question + Documents: Following the success of the retrieval augmented generation (RAG) approach in BioGen (Gupta et al., 2024), we devised another strategy in which the relevant documents, along with the question, were passed as input to the LLMs. To get the relevant documents, we used the two-stage approach, in which we first used BM25 to retrieve the top 100 relevant documents from the BioGen 2024 PubMed corpus, and then re-ranked and selected the top 10 relevant documents using GraphMonoT5 approach (Gupta and Demner-Fushman, 2024). We aimed to investigate the role of input documents in the model's capability of refining the final nuggets. Toward this, we developed two variants of this approach. In the first variant, we feed all the retrieved documents together to the model, and in the second variant, we feed each document sequentially and instruct the model to refine the nuggets and produce the final nuggets at the end of the iteration. We call the former variant QRD_{all} and the latter QRD_{seq}. We also extended this strategy to the ground-truth documents and used the cited documents associated with each assertion in the reference answer. We call these variants QGD_{all} (all documents together) and QGD_{seq} (sequential documents).

(4) Question + Answer + Documents: Similar to the Question + Documents strategy, we devise other strategies where we include the ground-truth answer in the sequential processing of documents (QRD_{seq}, QGD_{seq}) and all documents together (QRD_{all}, QGD_{all}) settings, and call them ($QARD_{seq}$, $QAGD_{seq}$) and ($QARD_{all}$, $QAGD_{all}$) for sequential processing of documents and all documents together, respectively.

We have provided all the prompts and experimental details in the **Appendix**.

2.3 Evaluation Metrics

For a given question Q, and its ground-truth nuggets $Y = \{y_1, y_2, \ldots, y_m\}$ and model nuggets $X = \{x_1, x_2, \ldots, x_n\}$ of the size m and n respectively, we aim to match each nugget $X_i \in X$ to one of the ground-truth nuggets $y_j \in Y$. We formulate the nuggets matching as an assignment problem, where we first compute the semantic similarity $sim(x_i, y_j) = cosine(emb_{x_i}, emb_{y_j})$ between $x_i \in X$ and $y_j \in Y$ and create a similarity matrix $S \in \mathcal{R}^{m \times n}$. We then group all the elements of matrix S and sort them in descending order. Iteratively, we assign each x_i to y_j if, $S_{ij} >= \theta$, and x_i and y_j have not been assigned. We continue the process until all x_i (having $sim(.) >= \theta$) has been assigned. We keep track of each assigned y_j and ensure each y_j is mapped to at most one x_i while maximizing global similarity. Once the assignment is done, we compute precision $p = \frac{|X \cap Y|}{m}$ and recall $r = \frac{|X \cap Y|}{n}$, and F1-score, where $|X \cap Y|$ denotes the number of generated nuggets that match the ground-truth nuggets. For computing semantic similarity, we use the SentenceTransformer (Reimers and Gurevych, 2019) model[2].

3 Results and Discussion

Key Results: Table 2 shows the experimental results of multiple generation strategies at the optimal[3] value of threshold θ. For the **Question** strategy, the Llama 3.3 (70B) model obtained the maximum F1-score of 34.03% in zero-shot setting. On the **Question + Answer** strategy, which can be considered as an upper-bound for the LLMs, the Llama 3.3 (70B) model achieved a maximum F1-score of 76% in one-shot setting. Under **Question + Document** strategy, all the LLMs exhibited suboptimal performance and showed a maximum of 34.11% F1-score with the Gemini 2.0 Flash model, where all the relevant documents along with the question are provided as input to the model. On the **Question + Answer + Documents** strategy, the Qwen 2.5 (72B) model achieved a maximum F1-score of 62.95% where the relevant documents (one at a time until all the documents finished), along with

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3]The optimal value (0.7) was determined by manual comparison of 10 different sets of LLM and ground-truth nuggets.

Models	Question		Question + Answer		Question + Documents				Question + Answer + Documents			
	Q_0	Q_1	QA_0	QA_1	QGD_{all}	QRD_{all}	QGD_{seq}	QRD_{seq}	$QAGD_{all}$	$QARD_{all}$	$QAGD_{seq}$	$QARD_{seq}$
DeepSeek-R1 (7B)	17.28	19.11	34.07	53.61	9.86	7.95	16.91	5.82	9.68	8.85	13.3	5.61
DeepSeek-R1 (70B)	25.56	28.39	62.97	68.1	14.3	8.55	19.53	13.82	10.82	10.87	33.9	28.36
Falcon 3 (7B)	25.67	24.54	54.42	44.83	13.16	9.37	25.24	17.21	14.52	11.56	45.36	43.23
Falcon 3 (10B)	23.5	28.21	51.76	60.32	13.69	9.87	21.73	13.22	12.41	9.01	47.61	48.76
Gemma 2 (9B)	23.91	23.93	50.4	62.03	11.67	10.95	19.07	17.95	13.1	11.21	43.38	37.19
Gemma 2 (27B)	27.11	27.42	59.61	65.16	14.0	9.52	18.42	13.09	13.03	13.32	34.15	31.39
Llama 3.2 (3B)	19.15	15.52	37.14	50.32	13.6	11.56	18.36	11.37	15.17	8.82	39.52	41.57
Mistral Small (24B)	26.97	28.65	41.02	67.41	12.56	9.04	21.25	17.16	13.62	8.34	38.53	32.85
Phi-4 (14B)	26.57	26.32	61.84	66.33	12.68	9.23	26.29	15.68	12.4	10.98	39.03	37.11
Qwen2.5 (7B)	22.43	25.96	64.38	65.2	11.24	7.86	20.43	12.56	11.21	9.13	34.11	25.69
Qwen2.5 (72B)	28.39	34.52	67.45	72.68	10.96	8.64	29.34	**24.39**	12.3	9.31	**56.41**	**62.95**
Vicuna1.5 (7B)	17.71	21.13	53.54	41.12	7.3	4.27	13.43	12.4	7.72	6.32	36.31	36.33
Vicuna1.3 (33B)	18.48	23.15	55.24	60.63	8.15	5.07	15.93	9.61	7.33	6.32	32.53	23.08
Llama 3.3 (70B)	**34.03**	**33.45**	**68.32**	**76**	17.53	10.94	**29.37**	22	18.29	12.72	39.1	44.8
GPT-4o	33.48	31.22	64.03	69.82	29.63	**20.16**	24.55	21.17	7.63	10.2	24.39	44.17
Gemini 2.0 Flash	33.14	35.32	56.05	72.55	**34.11**	15.29	8.81	19.85	50.83	43.68	32.64	31.02
Claude 3.5 Sonnet	27.08	16.3	66.11	67.86	33.17	17.44	17.16	16.56	**51.48**	**43.82**	47.87	46.18

Table 2: Performance comparison of various open and closed-source LLMs on the task of nugget generation under different generation strategies. All the results are reported here denote the F1-score.

the question and answer, are provided as input to the model.

Discussion and Findings: We observed a **significant performance gap among the generation strategies**. The best model's performance difference between **Question + Answer** and **Question** is 41.97%. Similarly, we recorded a performance difference between **Question + Answer** and **Question + Document** as 41.89%. With the ground-truth documents as well, the GPT-4o obtained an F1-score of 29.63% compared to its counterpart **Question** strategy with an F1-score of 33.48%. Similar observations are made for most of the open-source LLMs, except for the Gemini 2.0 Flash model, where the difference between **Question + Documents** (34.11) and **Question** (33.14) strategy is not significant.

We observed that smaller models (3B-14B) tend to obtain lower performance compared to their counterpart larger models. The study also reveals that **LLMs lack the capability of accurately generating or extracting nuggets** for the health-related **query**. Table 2 exhibits the performance of **Question** strategy that tests the LLMs' knowledge in generating the nuggets for the given question, which does not achieve the anticipated performance. For the **Question + Answer** strategy, where the ground-truth answer was given to the model to extract the nuggets, it only achieves the F1-score of 76% which highlights LLMs limitation in accurately distilling the atomic facts from the answer. LLMs showed similar behavior when the documents were given to the model for generating/extracting the nuggets.

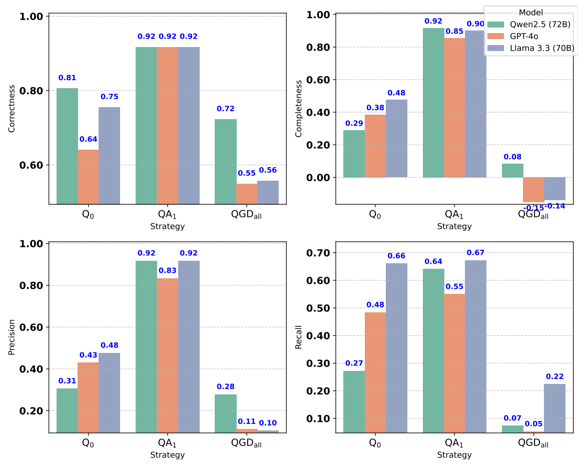

Figure 1: Performance comparison of different models under different generation strategies on multiple human evaluation criteria.

Human Evaluation: We also performed an extensive human analysis on the LLM-generated nuggets on multiple evaluation criteria. For the human evaluation, we chose the top-3 best-performing LLMs across multiple settings. We evaluated a total of 320 model-generated nuggets with 37 ground-truth nuggets for three diverse questions across three different settings: Q_0, QA_1, and QGD_{all}. We evaluate the quality of each nugget on the following criteria: **(a) Correctness**: whether the generated nugget is correct (2), partially correct (1), and incorrect (0); **(b) Completeness**: whether the generated nugget is misleading (-1), not required (0), Okay, but not required (1), and required (2); **(c) Precision**: portions of the generated nuggets that are correct; **(d) Recall**: portions of the ground-truth nuggets covered in the generated nuggets. We computed all the aforementioned scores for each question and averaged them to report (Fig. 1) the overall scores. The human evaluation of completeness and correctness criteria reveals that under the ground-truth answers (QA_1 strategy) all three LLMs' performance was better, only the question (Q_0) strategy obtained the sub-optimal performance. The evaluation also highlights that automatic precision and recall are highly aligned with manual precision and recall.

4 Conclusions

This work presented a comprehensive study on generating nuggets for health-related questions using various open and closed-source LLMs. Firstly, we manually formulated nuggets for BioGen 2024 topics and thereafter, we devised multiple nugget generation strategies to assess the capability of LLMs under different settings. We found that most LLMs obtained sub-optimal performance on the task which demonstrates the challenge involved with nugget generation, and we believe that our manual-created nuggets will promote further research in this direction.

Ethics Statement

The health-related questions and reference answers used in this study are publicly available within the Text REtrieval Conference (TREC) 2024 data.

Limitations

While the BioGen 2024 dataset covers a broad range of question topics and intents sourced out of popular health-related searches, it is not exhaustive. Subsequently, our findings on the LLMs' ability to generate nuggets in zero-shot settings apply to information needs covered in the data: clinical decision-support, factoid, and treatment and environment effects. The manual nugget evaluation approach outlined above can be used in the future to expand the data.

Another limitation of the data is a single reference answer. While the bulk of the nuggets must be present in any answer, some of the automatically generated nuggets could have been present in alternative answers. For example, if the existing reference answer list surgery as a treatment option, without specifying the best procedures, automatically generated nuggets that name specific surgeries will not get any credits for these nuggets. While this may somewhat lower the scores, it should not affect the model ranking, as the same approach is used for all models. In the future, more than one reference answer would be desirable to base the evaluation on a nugget pyramid (Marton and Radul, 2006).

Acknowledgments

This research was supported by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai

Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Laura Dietz. 2024. A workbench for autograding retrieve/generate systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1963–1972, New York, NY, USA. Association for Computing Machinery.

Naghmeh Farzi and Laura Dietz. 2024. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, page 175–184, New York, NY, USA. Association for Computing Machinery.

Travis R Goodwin, Dina Demner-Fushman, Kyle Lo, Lucy Lu Wang, Hoa T Dang, and Ian M Soboroff. 2022. Automatic question answering for multiple stakeholders, the epidemic question answering dataset. *Scientific Data*, 9(1):432.

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. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Deepak Gupta and Dina Demner-Fushman. 2024. Empowering language model with guided knowledge fusion for biomedical document re-ranking. In *International Conference on Artificial Intelligence in Medicine*, pages 251–260. Springer.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *Preprint*, arXiv:2411.18069.

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.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T

McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Gregory A. Marton and Alexey Radul. 2006. Nuggeteer: Automatic nugget-based evaluation using descriptions and judgements. In *North American Chapter of the 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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework. *arXiv preprint arXiv:2411.09607*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, 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, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. 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-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, Lena Heuermann, Leti cia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, 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, Peng chong 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, S'ebastien M. R. Arnold, Se bastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, 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, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément 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. *ArXiv*, abs/2408.00118.

Kensen Shi, Deniz Altınbüken, Saswat Anand, Mihai Christodorescu, Katja Grünwedel, Alexa Koenings, Sai Naidu, Anurag Pathak, Marc Rasi, Fredde Ribeiro, et al. 2024. Natural language outlines for code: Literate programming in the llm era. *arXiv preprint arXiv:2408.04820*.

Ian Soboroff. 2024. Don't use llms to make relevance judgments. *Preprint*, arXiv:2409.15133.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ellen M. Voorhees. 2004. Overview of the trec 2003 question answering track. In *Text Retrieval Conference*.

Huimin Wang, Yutian Zhao, Xian Wu, and Yefeng Zheng. 2024. imapScore: Medical fact evaluation made easy. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10242–10257, Bangkok, Thailand. Association for Computational Linguistics.

Phoebe J Wang and Max Kreminski. 2024. Guiding and diversifying llm-based story generation via answer set programming. *arXiv preprint arXiv:2406.00554*.

A Models

Table 3 is an exhaustive list of the models tested in our experiments with their versions and approximate number of parameters.

Model	Version / Size
Llama 3.2	3B[4]
Llama 3.3	70B[5]
Gemma 2	9B[6], 27B[7]
Mistral Small	24B[8]
Phi-4	14B[9]
Qwen2.5	7B[10], 72B[11]
$\text{Vicuna}_{1.5}$	7B[12]
$\text{Vicuna}_{1.3}$	33B[13]
Falcon 3	7B[14], 10B[15]
DeepSeek-R1	7B[16], 70B[17]
GPT-4o	`gpt-4o-2024-08-06`[18]
Claude 3.5 Sonnet	`claude-3-5-sonnet-20240620`[19]
Gemini 2.0 Flash	`gemini-2.0-flash`[20]

Table 3: A list of the models tested in our experiments.

B Details of Experimental Setup

Details of LLM Prompting To test the capabilities of LLMs to generate and extract nuggets, we prompted each model with a series of approaches. Table 4 contains the prompt used for Q_0, the zero-shot variation of our **Question** strategy. Table 5 contains the prompt used for QA_0, the zero-shot variation of our **Question + Answer** strategy. Table 6 contains the prompt used for Q_1, the one-shot variation of our **Question** strategy. Table 7 contains the prompt for QA_1, the one-shot variation of our **Question + Answer** strategy. Table 8 contains the prompt for QGD_{all} and QRD_{all}, our **Question + Document** strategies with all ground-truth documents and all retrieved documents, respectively. Table 9 contains the prompt for QRD_{seq} and QGD_{seq}, our **Question + Document** strategies with sequential ground-truth documents and sequentially retrieved documents, respectively. Table 10 contains the prompt for $QAGD_{all}$ and $QARD_{all}$, our **Question + Answer + Document** strategies with all ground-truth documents and all retrieved documents, respectively. Table 11 contains the prompt for $QARD_{seq}$ and $QAGD_{seq}$, our **Question + Answer + Document** strategies with sequential ground-truth documents and sequentially retrieved documents, respectively.

Each prompt contains some combination of query (q), answer (a), context (c), and initial nugget list (i) variables. The query and answer variables were substituted with each query and answer from the BioGen 2024 dataset. For the settings with all documents, the context variable was substituted with a list of the abstracts from all documents for the query separated by a new line character. For the settings with sequential documents, the context variable was substituted with a single abstract. The prompt for the sequential documents settings also contained the initial nugget list variable. This variable was initially "None" and then was substituted with the list of nuggets produced by the model provided with the previous abstract. For the sequential documents settings, the models were prompted once with each abstract and only the final list of nuggets was recorded. All models were prompted with their default settings (e.g. temperature).

[4] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[5] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[6] https://huggingface.co/google/gemma-2-9b-it
[7] https://huggingface.co/google/gemma-2-27b-it
[8] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501
[9] https://huggingface.co/microsoft/phi-4
[10] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[11] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[12] https://huggingface.co/lmsys/vicuna-7b-v1.5
[13] https://huggingface.co/lmsys/vicuna-33b-v1.3
[14] https://huggingface.co/tiiuae/Falcon3-7B-Instruct
[15] https://huggingface.co/tiiuae/Falcon3-10B-Instruct
[16] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[17] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B
[18] https://openai.com/index/hello-gpt-4o/
[19] https://www.anthropic.com/claude/sonnet
[20] https://deepmind.google/models/gemini/flash/

SYSTEM: You are NuggetGenerateLLM, an AI assistant specialized in generating all information nuggets that are required to completely answer a given query. A nugget is an atomic fact. **USER:** Generate all the information nuggets that are required to completely answer the query given below. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format: nugget1 nugget2 ... Query: q **LLM:** nugget1 nugget2 ...

Table 4: Prompt for Q_0.

SYSTEM: You are NuggetExtractLLM, an AI assistant specialized in extracting information nuggets from a given answer. A nugget is an atomic fact. **USER:** Generate all the information nuggets that are required to completely answer the query given below. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format: nugget1 nugget2 ... Query: q Answer: a **LLM:** nugget1 nugget2 ...

Table 5: Prompt for QA_0.

SYSTEM: You are NuggetGenerateLLM, an AI assistant specialized in generating all information nuggets that are required to completely answer a given query. A nugget is an atomic fact. **USER:** Generate all the information nuggets that are required to completely answer the query given below. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format: nugget1 nugget2 ... Here is an example query: Why is transferrin and iron low in covid patients but ferritin high? This is the list of nuggets that should be generated for this query: Lymphocytes and viruses compete for iron. Lymphocytes need iron for cellular response. Lymphocytes need iron for humoral response. Viruses need iron to replicate. Infection lowers iron levels in the blood. Infection increases ferritin levels in the blood. High ferritin is associated with increased mortality. Iron homeostasis needs ferritin. Ferritin is involved in physiologic processes. Ferritin is involved in pathologic processes. High ferritin indicates response to inflammation. High ferritin levels are linked to poor outcomes of COVID-19. Iron depletion therapy showed anti-viral activity in the COVID-19 pandemic. Iron depletion therapy showed anti-fibrotic activity in the COVID-19 pandemic. Query: q **LLM:** nugget1 nugget2 ...

Table 6: Prompt for Q_1.

SYSTEM: You are NuggetExtractLLM, an AI assistant specialized in extracting information nuggets from a given answer. A nugget is an atomic fact. **USER:** List all of the information nuggets in the answer given below that are required to completely answer the query. Each nugget must contain one, and only one, fact from the answer. A nugget must be as concise and as specific as possible. Each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format: nugget1 nugget2 . . . Here is an example query: Why is transferrin and iron low in covid patients but ferritin high? This is the list of nuggets that should be generated for this query: Lymphocytes and viruses compete for iron. Lymphocytes need iron for cellular response. Lymphocytes need iron for humoral response. Viruses need iron to replicate. Infection lowers iron levels in the blood. Infection increases ferritin levels in the blood. High ferritin is associated with increased mortality. Iron homeostasis needs ferritin. Ferritin is involved in physiologic processes. Ferritin is involved in pathologic processes. High ferritin indicates response to inflammation. High ferritin levels are linked to poor outcomes of COVID-19. Iron depletion therapy showed anti-viral activity in the COVID-19 pandemic. Iron depletion therapy showed anti-fibrotic activity in the COVID-19 pandemic. Query: q Answer: a **LLM:** nugget1 nugget2 . . .

Table 7: Prompt for QA_1.

SYSTEM: You are NuggetGenerateLLM, an AI assistant specialized in using context to generate all information nuggets that are required to completely answer a given query. A nugget is an atomic fact.
USER: Use the context provided to generate all the information nuggets that are required to completely answer the query given below. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format:
nugget1
nugget2
...
Query: q
Context: c
LLM:
nugget1
nugget2
...

Table 8: Prompt for QGD_{all} and QRD_{all}.

SYSTEM: You are NuggetGenerateLLM, an AI assistant specialized in using context to update a list of all information nuggets that are required to completely answer a given query. A nugget is an atomic fact.
USER: Use the context provided to update the list of information nuggets, if needed. The list should contain all nuggets that are required to completely answer the query given below. If no list of nuggets is provided, generate a list of nuggets. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format:
nugget1
nugget2
...
Query: q
Context: c
Initial Nugget List: i
LLM:
nugget1
nugget2
...

Table 9: Prompt for QGD_{seq} and QRD_{seq}.

SYSTEM: You are NuggetGenerateLLM, an AI assistant specialized in using context to generate all information nuggets that are required to completely answer a given query. A nugget is an atomic fact.
USER: Use the context provided to generate all the information nuggets that are required to completely answer the query given below. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format:
nugget1
nugget2
...
Query: q
Answer: a
Context: c
LLM:
nugget1
nugget2
...

Table 10: Prompt for $QAGD_{all}$ and $QARD_{all}$.

SYSTEM: You are NuggetGenerateLLM, an AI assistant specialized in using context to update a list of all information nuggets that are required to completely answer a given query. A nugget is an atomic fact.
USER: Use the context provided to update the list of information nuggets, if needed. The list should contain all nuggets that are required to completely answer the query given below. If no list of nuggets is provided, generate a list of nuggets. Each nugget must contain one, and only one, fact. A nugget must be as concise and as specific as possible. A nugget cannot contain a list, each element in a list must be its own nugget. Each nugget must directly answer the query. The list of nuggets must not contain redundant information. Return a list of nuggets such that each nugget is on a new line. Do not number or bullet the list. Do not include anything in your response except for the list of nuggets. Here is an example of the output format:
nugget1
nugget2
...
Query: q
Answer: a
Context: c
Initial Nugget List: i
LLM:
nugget1
nugget2
...

Table 11: Prompt for $QAGD_{seq}$ and $QARD_{seq}$.