# ChemLit-QA: A human evaluated dataset for chemistry RAG tasks

**Geemi P. Wellawatte**[1‡] **Huixuan Guo**[123‡] **Magdalena Lederbauer**[134]
**Anna Borisova**[13] **Matthew Hart**[135] **Marta Bucka**[13] **Philippe Schwaller**[13]
[1] Laboratory of Artificial Chemical Intelligence (LIAC), ISIC, EPFL
[2] Department of Chemical and Biomolecular Engineering, NTU Singapore
[3]National Centre of Competence in Research (NCCR) Catalysis, EPFL
[4] Department of Chemistry and Applied Biosciences, ETH Zurich
[5] Department of Applied Physical Sciences, UNC Chapel Hill
[‡] Contributed equally.
{geemi.wellawatte,philippe.schwaller}@epfl.ch

## Abstract

The evaluation of Large-Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems, particularly in knowledge-intensive fields like chemistry, is hindered by the number of available high-quality datasets. Existing datasets are often small due to the labor-intensive nature of manual curation, or require further quality checks when generated automatically. This study addresses the need for robust scientific datasets by introducing ChemLit-QA, an open-source, expert-validated dataset with over 1,000 entries tailored for chemistry. The dataset was initially generated using an automated framework and subsequently reviewed by experts.

## 1 Introduction

Over the last couple of years, we have seen an increased interest in developing and using Large Language Models (LLMs) to accelerate scientific discovery.[1–4] Their capability to compile vast amounts of information has opened up new avenues for knowledge processing and extraction.[1,5–7] While LLMs perform well in information retrieval and certain creative tasks within general knowledge domains, they often fall short in specialized, knowledge-intensive scientific fields such as chemistry.[8,9] In a similar study, Wang et al.[10] evaluated the performance of LLMs in materials science and highlighted the limitations of LLMs in domain-specific tasks. Often, when LLMs are prompted with queries in specialized domains where pre-training data is limited, they are prone to hallucinations and may generate false information.[11–14] Widely adopted strategies to address these limitations involves Retrieval-Augmented Generation (RAG)[15], fine-tuning[3,16,17], in-context learning[18,19] and, Language-Interfaced Fine-Tuning (LIFT)[20,21]. However, the success of these approaches relies on the availability of curated high-quality, task-specific datasets for training and benchmarking.

The most common RAG-specific datasets are in the form of Question-Answer (QA) pairs or Question-Answer-Context (QAC) triplets. The answers could be open-domain, yes-no type, or multiple-choice type. Creating high-quality, QA-type datasets poses significant challenges, especially in scientific fields. Manual curation by domain experts, while producing high-quality data, is resource-intensive. while automated generation using LLMs offers a potential solution. However, this approach poses the risk of hallucinations, propagating inaccuracies and biases present in pre-training data.[22–25] Other limitations include context limitations where QA pairs are generated without corresponding text[26], uneven information distribution, limited reasoning complexity, and lack of (RAG) specificity.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

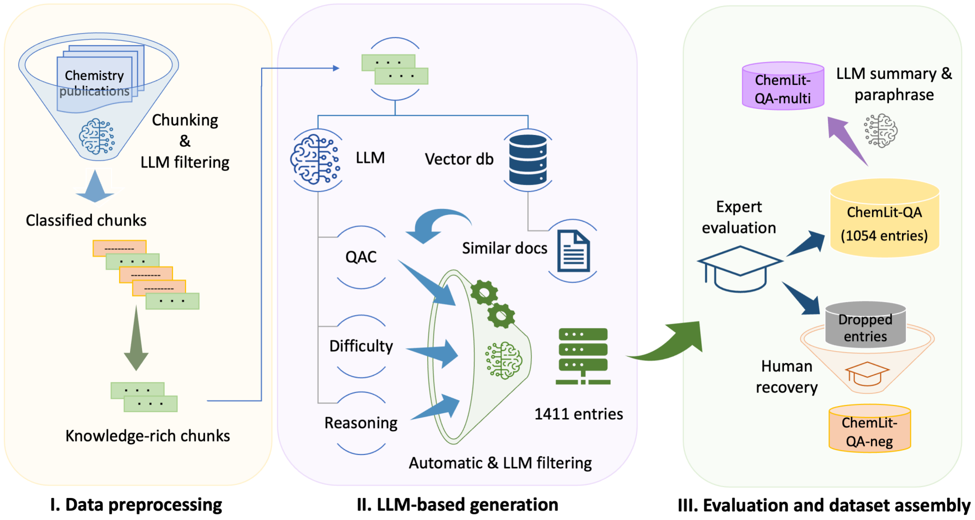


Figure 1: Dataset generation and evaluation pipeline used in this work.

To address these challenges, we present ChemLit-QA, an open-source, expert-validated, large dataset with more than 1000 entries, designed specifically for RAG and fine-tuning benchmark tasks in chemistry. We evaluate the efficacy of ChemLit-QA through downstream experiments (RAG and fine-tuning) with state-of-the-art LLMs, demonstrating its utility in generating knowledge-intensive, context-specific, human-like QAC triplets. Our results show improved performance of fine-tuned LLMs in RAG settings. Additionally, we present two additional datasets, 1) ChemLit-QA-neg: a dataset of 139 entries aimed to help detect hallucinations with questions where answers are not available, and 2) ChemLit-QA-multi: a dataset with 742 entries for multihop reasoning tasks. We highlight areas for further improvement, particularly in negative response identification tasks. Our datasets and source code are publicly available at `https://github.com/geemi725/ChemLit-QA`.

## 2 Method: Dataset curation

**1. Literature corpus** Downloaded ChemRxiv papers (`https://chemrxiv.org/`), up to March 2024 using "paperscraper" (`https://pypi.org/project/paperscraper/`) in PDF format and parsed to TEI XML format using Grobid (`https://grobid.readthedocs.io/en/latest`). A detailed procedure can be found in the Grobid GitHub repository. ChemRxiv was chosen as our data source because it is an open-access preprint server specifically for chemistry articles, making it ideal for building a chemistry-focused RAG benchmarking dataset.

**2. Paper classification and sampling** Next, we classified the parsed papers using a two-level hierarchy. Mistral-7B[27] model was used to assign labels based on title and abstract. Five papers were randomly selected from each second-level category. This clustering step aimed to enforce diversity within the generated dataset.

**3. Text chunking and usefulness identification** We used LangChain(`https://www.langchain.com/`) to construct our pipeline. The context of each paper was split recursively into chunks of a maximum length 2,000 characters. Next, each text chunk (split) was classified based on its content – chunks were retained only if they contained substantive scientific information such as experimental methods, results, theoretical concepts, chemical reactions, or technical discussions. Chunks were discarded if they contained only non-technical content such as author names, references, acknowledgments, funding information, or general background statements without specific scientific claims. For example, a chunk containing "This work was supported by NSF grant..." would be discarded, while one describing “The reaction yielded 85% conversion at 75°C..." would be retained. We employed GPT-4 Turbo[28] in this task.

**4. QA generation and Reasoning type classification** Again, we used GPT-4 Turbo[28] to generate questions from one of the seven types of reasoning – Explanatory, Comparative, Conditional, Causal, Predictive, Procedural, and Evaluative, to ensure their diversity and reasoning-intensiveness. Given a context, specially prompted LLM chains were used to generate QA pairs based on the reasoning types. Figure 2 panel (c) shows an example entry of the final dataset.

**5. Difficulty assignment** Next, we used LLM classifiers (GPT-4 Turbo[28]) to label the question-answer difficulty as ’Easy’, ’Medium’, or ’Hard’. For example, a question is easy to answer if the answer is directly available in a single sentence of the chunk.

**6. Similar chunk identification and start-end indices for answers** We used FAISS vector database (`https://faiss.ai/index.html`) to identify the top 4 similar chunks to a given context in a QAC triplet within the same document. `similarity_search_by_vector` method in FAISS uses Euclidean (L2) similarity for this. We employed SpaCy(`https://spacy.io/`) and Python scripting to determine start-end indices of answers within contexts. The goal of these similar chunks are to a) simulate a basic retrieval process and b) enhance benchmarking capabilities.

**7. Automatic filtering of generated data** After running the full QAC generation pipeline, we randmly sampled 2,000 entries and applied 4 LLM-based metrics (using DeepEval framework (`https://github.com/confident-ai/deepeval`) with GPT-4o) to validate the entries. The metrics are: answer relevancy, answer faithfulness, hallucination, and question faithfulness (customized with G-Eval) all in the range $\{0, 1\}$. Finally, we removed the QAC triplets according to the following conditions: a) question/answer faithfulness and answer relevancy scores is less $<$ mean $- 0.5 \times$ SD, b) hallucination score $> 0.1$, c) pSE $> Q_3 + 1.5 \times I_{QR}$ ($I_{QR}$: quartiles $Q_3 - Q_1$) This process resulted in a filtered dataset with 1,411 entries (dropped 589 entries).

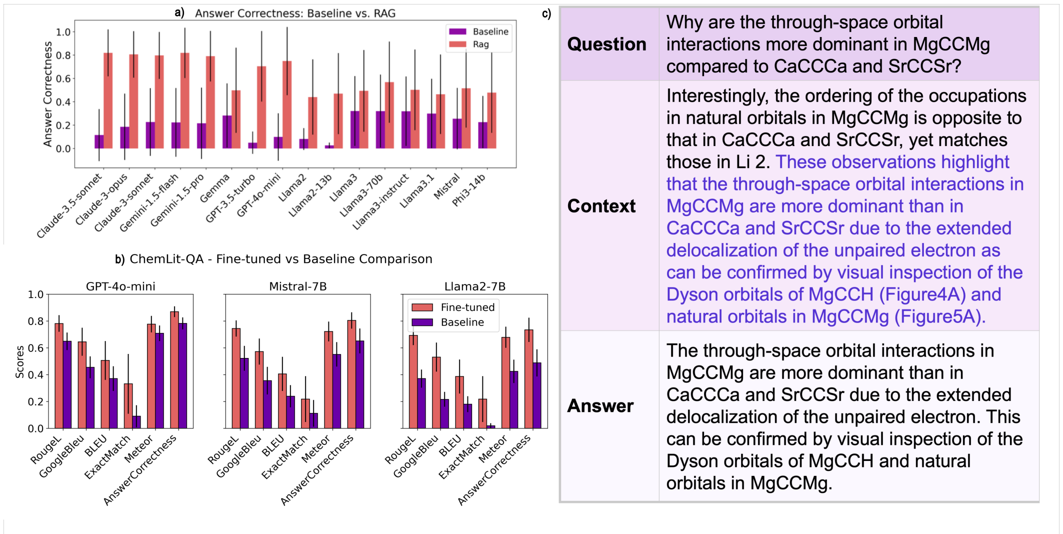
**8. Expert evaluation of the dataset** Once the dataset was filtered after step 7, we further evaluated the dataset using 4 human experts; 1 post-doctoral researcher, 2 PhD students, and 1 Bachelor’s student in Chemistry and Chemical engineering. Each evaluator was assigned an equal split of enries from the dataset and the entry was removed if it was marked as “drop”. An in-house interface was used for this purpose. This resulted in a further removal of 357 entries. Hence, the final ChemLit-QA dataset contained **1054 entries**. However, we also assessed inter-evaluator agreement, an additional subset of 60 entries was evaluated by all experts. Agreement levels were defined as: a) *Completely Agree* All 4 experts agree, b) *Almost agree* 3 out of 4 agree, c) *Partially agree* 50-50 agreement, and d) *Disagree* Any other case. This process ensured rigorous human validation of the dataset, enhancing its quality and reliability for benchmarking purposes.

## 3 Results

**Diversity of questions** ChemLit-QA contains 39.8% – Explanatory, 25.1% – Causal, 17.7% – Comparative, 7.4% – Procedural, 3.7% – Conditional, 3.4% –Evaluative and 2.8% – Predictive questions. On the other hand, it contained 74%, 18%, and 8% easy, medium, and hard questions in the final dataset respectively. We find that ChemLit-QA remains diverse in topics, covering 7 high-level and 12 low-level topics in chemistry, pre-defined by experts. The highest percentage of 32.25% of entries belonged to *quantum and theoretical chemistry*. The diversity of ChemLit-QA is also featured in the wide range of question keywords and phrases. Besides wh-determiners such as “why”, “what”, a considerable proportion of questions contain context-specific keywords that appear less than 4 times in the entire dataset, which are classified under “OTHERS”.

**Human and model agreements** We analyzed how the LLM-based metrics correlate with the human classification from the expert-evaluation results. Here we observed that the LLM and humans agree on 76%, 61%, and 6% for easy, medium, and hard questions respectively. Based inter-evaluator agreement results, we see that 84%-95% times humans either completely or substantially agree with each other.

**RAG: benchmarking LLMs on ChemLit-QA** We benchmarked different LLMs in a RAG pipeline using the ChemLit-QA dataset (QAC triplets + similar chunks, tested on a random subset of 211 entries). As shown in Figure 2 a), our results agree with the previous findings that a RAG approach is better at generating factual answers than relying on the model's internal knowledge.[15,29–31] All 16 models demonstrated a significant performance increase when provided with contextual information.



| | |
|---|---|
| **Question** | Why are the through-space orbital interactions more dominant in MgCCMg compared to CaCCCa and SrCCSr? |
| **Context** | Interestingly, the ordering of the occupations in natural orbitals in MgCCMg is opposite to that in CaCCCa and SrCCSr, yet matches those in Li 2. These observations highlight that the through-space orbital interactions in MgCCMg are more dominant than in CaCCCa and SrCCSr due to the extended delocalization of the unpaired electron as can be confirmed by visual inspection of the Dyson orbitals of MgCCH (Figure4A) and natural orbitals in MgCCMg (Figure5A). |
| **Answer** | The through-space orbital interactions in MgCCMg are more dominant than in CaCCCa and SrCCSr due to the extended delocalization of the unpaired electron. This can be confirmed by visual inspection of the Dyson orbitals of MgCCH and natural orbitals in MgCCMg. |

Figure 2: a) Mean Answer correctness scores of RAG and baseline models. b) Comparison between baseline and fine-tuned performance on the test dataset for GPT-4o mini[32], Mistral-7B[27], and Llama2-7B[33]. Evaluated on a test dataset with 211 entries. Error bars represent the standard deviation of the scores.c) Sample entry from ChemLit-QA. Evidence sentence used to answer the question is highlighted in purple

.

**Fine-tuning on ChemLit-QA** We used the ChemLit-QA dataset to fine-tune 3 LLM models (Llama2-7B[33], Mistral-7B[27] and GPT-4o mini[32]), tested on the same 211 entries. While we acknowledge that more advanced models have since been released, our primary goal was to demonstrate that fine-tuning on domain-specific data improves model performance regardless of the base architecture. Results in Figure 2 b) show that our results agree with previous studies[34,35] that fine-tuning improves the inherent performance of an LLM. In particular, we see significant performance in Llama2 compared to the other 2 models, its Answer Correctness increases from an average of 0.49 to 0.73. We expect that applying the same fine-tuning approach to more recent, cutting-edge LLMs would yield even better results, but this demonstration effectively shows the value of our ChemLit-QA dataset for improving chemistry-specific performance.

**Evaluations on ChemLit-QA-neg & ChemLit-QA-multi** One of the limitations of LLMs is hallucinating answers, specifically when they cannot be deduced based on the given contexts. We tested this on our ChemLit-QA-neg dataset, where we evaluated the LLM's ability to identify negative questions by either responding "Answer not available from the context" or arriving at the same conclusion through reasoning. Based on these results, we see that only GPT-4o mini[32] and Claude-3.5-Sonnet[36] achieved an Answer Correctness score over 0.5, scoring 0.58 and 0.55, respectively. When we evaluated the same LLMs as Figure 2 a), in our ChemLit-QA-multi dataset, all models retained the same performance order but much lower in scores. This proves ChemLit-QA-multi dataset is more challenging than the main dataset.

## 4 Conclusion

We present ChemLit-QA, an open-source, expert-validated, QAC-type, dataset for RAG and fine-tuning tasks in chemistry. Additionally, we include semantically similar chunks in the dataset to simulate the retrieval component of an RAG pipeline. We benchmarked several state-of-the-art proprietary and open-source LLMs, finding that proprietary models outperformed open-source counterparts in RAG tasks. However, fine-tuning models like Mistral-7B[27] and Llama2-7B[33] yielded performance comparable to proprietary models. We also introduce two supplementary datasets, ChemLit-QA-neg for hallucination detection and ChemLit-QA-multi for reasoning-intensive tasks, revealing challenges that LLMs still face in achieving human-level intelligence. Limitations of our work include fixed-size text chunks and the need for multi-modal data extraction. Future research is aimed to explore these limitations. We anticipate that these datasets will significantly contribute to advancing scientific discovery.

## Acknowledgments

This publication was created as part of NCCR Catalysis (grant number 225147), a National Centre of Competence in Research funded by the Swiss National Science Foundation. GPW acknowledges funding from the EPFL large-scale Solutions4Sustainability demonstrator project (SusEcoCCUS).

## References

[1] Jaewoong Choi and Byungju Lee. Accelerating materials language processing with large language models. *Communications Materials*, 5(1):13, 2024.

[2] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[3] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[5] Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, et al. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376, 2023.

[6] Glen M Hocky and Andrew D White. Natural language processing models that automate programming will transform chemistry research and teaching. *Digital discovery*, 1(2):79–83, 2022.

[7] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelmi, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, Caroline T. Holick, Tanya Gupta, Mehrdad Asgari, Christina Glaubitz, Lea C. Klepsch, Yannik Köster, Jakob Meyer, Santiago Miret, Tim Hoffmann, Fabian Alexander Kreth, Michael Ringleb, Nicole Roesner, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. Are large language models superhuman chemists?, 2024.

[8] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.

[9] Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[10] Hongchen Wang, Kangming Li, Scott Ramsay, Yao Fehlis, Edward Kim, and Jason Hattrick-Simpers. Evaluating the performance and robustness of llms in materials science q&a and property predictions. *arXiv preprint arXiv:2409.14572*, 2024.

[11] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[12] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[13] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[14] Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

[15] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[16] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[17] Wei Zhang, Qinggong Wang, Xiangtai Kong, Jiacheng Xiong, Shengkun Ni, Duanhua Cao, Buying Niu, Mingan Chen, Yameng Li, Runze Zhang, et al. Fine-tuning large language models for chemical text mining. *Chemical Science*, 15(27):10600–10611, 2024.

[18] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[19] Tom B Brown. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.

[20] Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

[21] Mayk Caldas Ramos, Shane S Michtavy, Marc D Porosoff, and Andrew D White. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341*, 2023.

[22] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

[23] Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564, 2023.

[24] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[25] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

[26] Yuwei Wan, Aswathy Ajith, Yixuan Liu, Ke Lu, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. Sciqag: A framework for auto-generated scientific question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*, 2024.

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[28] Open AI. Gpt-4 turbo in the openai api, 2024.

[29] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[30] Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *arXiv preprint arXiv:2404.07220*, 2024.

[31] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

[32] Open AI. Gpt-4o mini: advancing cost-efficient intelligence, 2024.

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[34] Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. Enhancing large language model performance to answer questions and extract information more accurately, 2024.

[35] Jeyoon Yeom, Hakyung Lee, Hoyoon Byun, Yewon Kim, Jeongeun Byun, Yunjeong Choi, Sungjin Kim, and Kyungwoo Song. Tc-llama 2: fine-tuning llm for technology and commercialization applications. *Journal of Big Data*, 11(1):100, 2024.

[36] Anthropic. Claude 3, 2023.