HaloRAG: Towards Mitigating LLM Hallucinations with Low-Cost Real-Time Retrieval

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often struggle to stay up-to-date due to their reliance on static datasets, leading to outdated responses and hallucinations. We introduce HaloRAG, a cost-efficient agentic wrapper that enhances LLMs with real-time information retrieval using advanced web scraping technologies. Leveraging semantic searches and Retrieval-Augmented Generation (RAG), this wrapper fetches, validates, and summarizes up-to-date web data, extending the LLM's knowledge base without retraining. This method significantly enhances the accuracy and relevance of LLM responses, particularly for queries requiring the latest information. Comparative analysis indicates that the wrapper-enhanced LLM outperforms models like GPT-3.5 and Claude on queries involving recent events and emerging technologies. This work advocates for integrating real-time data retrieval techniques to significantly reduce hallucinations and extend the practical applicability of LLMs across various domains.

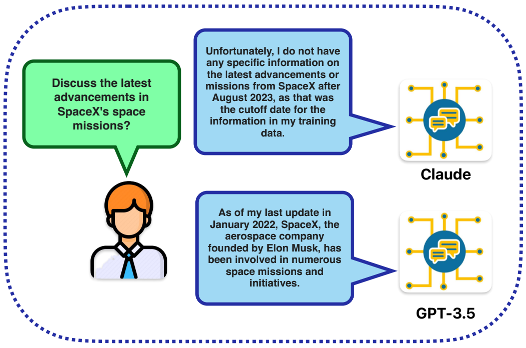

Figure 1: LLMs trained on older datasets are not capable of generating responses to queries involving recent recent developments.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing (NLP) with their ability to perform tasks such as text generation, translation, summarization, and question-answering. Models like GPT-3.5 by OpenAI (Brown et al., 2020) and Claude by Anthropic (Bai et al., 2022) demonstrate impressive capabilities. However, LLMs still face significant challenges, particularly the issue of hallucinations—generating information that appears plausible but is factually incorrect or outdated (Bender et al., 2021; Liu et al., 2021; Maynez et al., 2020; Ji et al., 2023). This issue often arises from limitations in their training datasets, which may not cover the latest information or be sufficiently comprehensive.

The lack of efficient mechanisms to integrate external data exacerbates these issues, as current methods frequently rely on expensive APIs (Thoppilan et al., 2022). Moreover, running LLMs locally can lead to high RAM consumption, making it impractical for users with limited computational resources. To address these limitations, we propose HaloRAG, a web-based Retrieval-Augmented Generation (RAG) (Lewis et al., 2021a) approach, which enhances the interaction between user queries and LLMs by integrating real-time information from the web. Our method significantly reduces hallucinations by grounding responses to updated information, thus enhancing accuracy and relevance by retrieving the most pertinent documents, and offers a cost-effective, scalable solution through automated web scraping techniques. By continually incorporating fresh data, HaloRAG ensures that the underlying LLMs remain current and capable of addressing queries related to recent developments and emerging topics, extending their utility across various dynamic informational contexts. This approach represents a low-cost step towards making LLMs more reliable, accurate, and applicable in real-time scenarios. To mitigate high resource consumption, the approach is designed to be efficient, allowing it to run on local computers with minimal resources and enabling users to benefit from enhanced LLM capabilities without requiring server infrastructure.

The key contribution of our work is a wrapper for LLMs with the following key features:

- Low computational overhead
- Ability to handle multilingual prompts
- Real-time data retrieval and integration
- Compatibility with various LLM architectures
- Reduced hallucinations in generated responses without the need for re-training with new data

2 Related Work

Hallucination in natural language generation has been identified as a major challenge, attributed to static training datasets that fail to reflect the evolving real-world knowledge (Ji et al., 2023). Dynamic updates to models are essential to mitigate this problem, with strategies that enhance model architectures and improve training data quality being crucial (Liu et al., 2023). Real-time data retrieval and the use of adversarial training are proposed to enhance factual accuracy in dialogue systems. Similar to our work, FreshLLMs (Vu et al., 2023) explores augmenting LLMs with web search capabilities. However, the approach diverges in critical ways. The FreshLLMs framework relies on the Google search API, which incurs costs and may not be accessible to all users. Moreover, FreshLLM is limited to English language prompts, whereas HaloRAG supports multilingual queries, offering a more versatile solution.

The limitations of static datasets in LLMs, which prevent access to up-to-date information and result in outdated responses, have been addressed by integrating real-time data sources to maintain model currency (Komeili et al., 2021). The Retrieval-Augmented Generation (RAG) approach has shown promise in improving LLM responses by using relevant documents to inform the generation process, which enhances both accuracy and contextual appropriateness (Lewis et al., 2021b).

Concerns about the dominance of the English language in LLMs have led to calls for the development of multilingual models to ensure inclusivity and accessibility (Bender et al., 2021). Cross-lingual model training has shown that LLMs can effectively operate across different languages, which is critical for global applicability (Lample and Conneau, 2019).

Finally, the high computational costs associated with large models highlight the need for efficient scaling strategies. These strategies should balance model size, computational efficiency, and performance to allow for scalable and cost-effective AI systems (Kaplan et al., 2020). The approach of refreshing LLMs with search engine augmentation presents a promising direction, although it currently faces challenges related to cost and language limitations (Vu et al., 2023).

3 Methodology

HaloRAG consists of five different modules and we describe each of it in this section. Figure 2 shows the architecture of our approach.

3.1 Multilingual Support

The inclusion of non-English speaking users is achieved through the integration of a multilingual support system. To this end, the user queries are first translated into the primary operating language of the LLM and subsequently the LLM's responses back into the user's language. Bidirectional translation ensures effective processing and response to queries from a diverse global audience. The component is designed to create a seamless interaction loop, thereby minimizing language barriers. This design enhances the usability and accessibility of LLM technologies across various linguistic demographics.

3.2 Integration with LLMs

The integration of large language models (LLMs) is designed to be straightforward and minimally invasive, allowing for the incorporation of various models without significant modifications to their architectures. This approach ensures broad applicability across different models and configurations, facilitating widespread adoption. The integration process supports any LLM, including those hosted on platforms such as HuggingFace, and is compatible with both general-purpose and specialized models.

For LLMs fine-tuned for specific applications, such as financial forecasting or medical research, this integration provides access to current external information, enhancing performance and accuracy. This allows LLMs to respond to queries based not only on their pre-trained knowledge but also on the

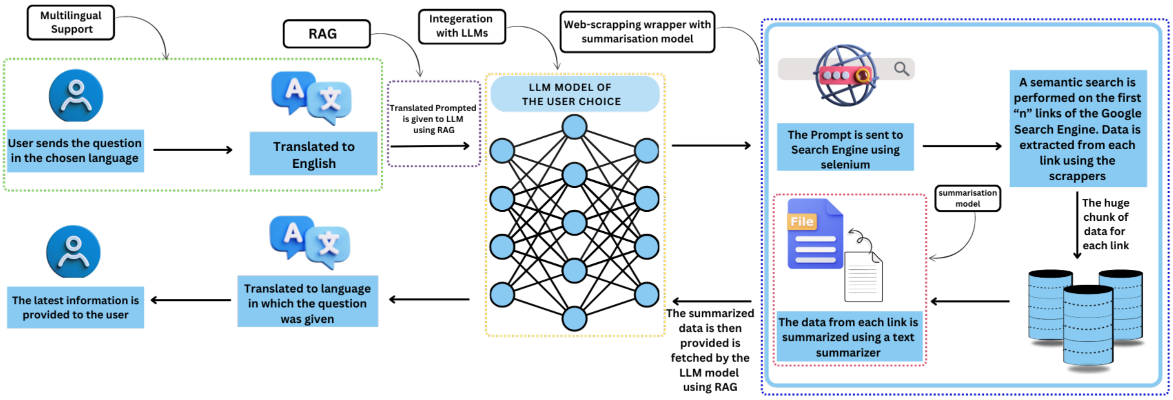

Figure 2: Architecture Diagram. Method for enhancing language model accuracy using web scraping and RAG. Starts with user queries in native languages,translation, web scraping, data summarisation, and re-integration into the LLM for refined responses.

latest available data. This capability is particularly valuable in dynamic and rapidly evolving information landscapes. The integration strategy interfaces with the model's input and output processes, enabling real-time data enhancement before response generation. This ensures that LLMs remain current and reliable, addressing the limitations of static training datasets and expanding the utility of LLMs across various applications and industries.

3.3 Retrieval-Augmented Generation (RAG)

The Retrieval-Augmented Generation (RAG) component is designed to enhance the interaction between a user's query and the LLM by integrating external data. This approach combines retrieval-based and generation-based methods to improve the accuracy and relevance of the LLM's responses (Lewis et al., 2021b). Upon receiving a user query, the LLM processes the input to understand its context and intent. The RAG component then formulates a search query to retrieve relevant information from various online sources. This retrieved data is evaluated for relevance and credibility, ensuring that only authoritative and reliable information is used. The integration of retrieval and generation in RAG allows the LLM to provide responses that are not only based on static knowledge but are also enriched with real-time, contextually appropriate information. This dynamic interaction significantly reduces the incidence of generating hallucinations—responses that are plausible but factually incorrect—by continually updating the LLM with the latest information (Izacard and Grave, 2021). By continually integrating fresh data from the web, the RAG component ensures that the LLM remains current and capable of addressing queries about recent developments and emerging topics.

3.4 Web Scraping Module

The web scraping module is designed to access recent and relevant information from the internet. Upon receiving a query, the module structures the input for search engines. The number of links processed is adjustable based on research depth and computational resource availability. The semantic analysis assesses the relevance of each link's metadata to ensure the accuracy and substance of the information. Each identified link is open in a controlled browser session, extracting textual information. Semantic processing is applied to ensure the extracted text is contextually relevant. The scraped data from each link is aggregated into unified data, undergoing preliminary processing to eliminate duplicates, correct formatting issues, and prepare the data. This method ensures the retrieval of the most current and relevant information, enhancing the system's overall effectiveness.

3.5 Summarization Models

The summarization models in the system are crucial for distilling extensive information retrieved from the web into concise, essential content that can be effectively utilized by large language models (LLMs). HaloRAG employs the FalconAI summarization model (Almazrouei et al., 2023) which is known for its efficiency and accuracy in processing large texts. FalconAI summarization model utilizes a trained network to extract key facts and themes, ensuring the summaries are both coherent and focused on the most relevant details. This process significantly reduces the informational load on LLMs, enabling quicker and more accurate responses. The system is designed with flexibility, allowing for the integration of other summarization models if required. This adaptability ensures continuous updating the approach to summarisation based on evolving technological advancements or specific application needs.

4 Experiments

To assess the performance of HaloRAG, a sample dataset consisting of factual prompts was curated. These prompts were specifically selected to cover recent developments (specifically from 2023 and 2024) in technology, politics, and related fields, hence allowing us to focus on the model's ability to handle up-to-date content. More details can be found in Appendix A. We compare the responses from several models, including GPT-3.5 (Brown et al., 2020), GPT-4o (OpenAI et al., 2024), Claude (Bai et al., 2022), Phind (Rozière et al., 2024), and HaloRAG with Qwen (Bai et al., 2023) as the underlying LLM. For our experiments we used Google Colab with a RAM of 12.5 GB and GPU of 15 GB.

5 Results

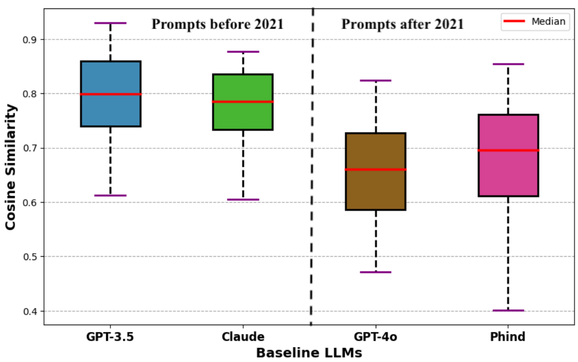

Figure 3: Cosine similarity between the responses generated by GPT-3.5, Claude, GPT-4o and Phind

Figure 3 shows the cosine similarity between the responses generated by our method (HaloRAG). The results show a median cosine similarity of 0.6599 with GPT-4o and 0.6956 with Phind, demonstrating textual and conceptual alignment between the responses generated. The highest cosine similarity recorded is 0.8240 with GPT-4o and 0.8545 with Phind whereas models like GPT-3.5 and Claude were unable to generate the responses for these same prompts (Fig 1). This shows that our model is able to generate responses to queries similar to those generated by state-of-the-art LLMs without the requirement of re-training with new data. In order to measure the effectiveness of HaloRAG with pre-2021 queries, we evaluate the cosine similarity between our approach and GPT-3.5 and Claude on another dataset consisting of 30 prompts centered around information before 2021 (more details can be found in Appendix B). The results indicate a median cosine similarity of 0.7997 with GPT-3.5 and 0.7848 with Claude, demonstrating textual and conceptual alignment between the responses generated.

6 Conclusion and Future Work

We presented HaloRAG , an approach to enhancing Large Language Models (LLMs) by developing a cost-efficient, no-cost agentic wrapper that leverages web scraping technologies to perform semantic searches and retrieve real-time information from the web. This wrapper utilizes RAG to extend the knowledge base of any LLM it is paired with, enabling it to provide accurate and current answers without needing to be retrained on new datasets. The approach demonstrates significant improvements over existing models like GPT-3.5 and Claude, particularly in handling prompts about recent events or emerging technologies. Future work includes improving the reliability and computational speed of the wrapper and performing a comprehensive comparison with other baseline LLMs on standard datasets.

7 Limitations

While the wrapper-enhanced LLM offers significant advancements, it is essential to acknowledge certain limitations. One notable limitation is the dependency on web scraping technologies, specifically Selenium and Beautiful Soup (bs4). If the structure of the Google search results page changes, the web scraping components may fail to function correctly, leading to incomplete or inaccurate data extraction. This reliance necessitates frequent updates to the scraping logic to adapt to changes in the search engine's HTML structure. The wrapper ensures that information is retrieved from verified sources, there remains a risk of encountering biased or misleading information. Users must critically evaluate the responses and cross-reference with other sources when necessary. In conclusion, while the wrapper-enhanced LLM reduces hallucinations and provides up-to-date information, it has limitations such as dependency on web scraping.

References

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? pages 610–623.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *Preprint*, arXiv:2007.01282.

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 Computing Surveys*, 55(12):1–38.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *Preprint*, arXiv:2107.07566.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Preprint*, arXiv:1901.07291.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021a. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Preprint*, arXiv:2101.06804.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. 2023. Understanding the difficulty of training transformers. *Preprint*, arXiv:2004.08249.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik

Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *Preprint*, arXiv:2201.08239.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation. *Preprint*, arXiv:2310.03214.

A Prompts based on information in 2023-2024

- Tell me about the current status of Palestine war?
- Discuss the latest developments in the Ukraine conflict?
- What are the current issues surrounding the U.S. immigration policy?
- Explain the controversy over the 2024 U.S. Presidential election.
- What are the recent advancements in renewable energy technology?
- What is the current status of the COVID-19 pandemic worldwide?
- Explain the controversy surrounding the recent tech layoffs?
- Discuss the impact of the recent stock market fluctuations due to recession?
- What are the latest trends in cryptocurrency regulations in 2024?
- Explain the current issues in global supply chain disruptions?
- Discuss the latest findings in climate change research from 2023 onwards?
- Explain the controversy over the new AI regulations in the EU since 2024?
- Discuss the recent developments in space exploration by India since 2023?
- Explain the impact of the recent data breaches occurred in AIMS India?
- Discuss the current debates on healthcare reform in the U.S. from January 2024?
- What is the current status of the trade war between the U.S. and China?
- Explain the controversy surrounding the recent Supreme Court decisions about the farmers protest?
- Explain the controversy revolving around NVIDIA and Jensen Huang in 2024?
- What are the latest advancements in cancer research?
- Explain the current issues in the global refugee crisis?
- Discuss the implications of the latest tech company mergers?
- Discuss the controversy of Google using Reddit for search AI?
- Tell me about the fight going between Elon and Mark?
- Tell me about the fight going between Elon and Yann LeCun?
- Explain the impact recent cyclone that hit the east coast of India?
- Discuss the current debates on WhatsApp shutting down its service in India?
- What is the current status of the opioid crisis in the U.S.?
- Discuss the latest advancements in SpaceX's space missions?
- Discuss the current debates on gun control laws in the U.S.?
- Tell me about the controversy of Devin-AI?

B Prompts based on information before 2021

- What were the key elements of the Paris Climate Agreement?
- Describe the impact of Brexit on the European Union's trade policies.
- How did the US-China trade war begin, and what were its major impacts?
- What are the principles of the Green New Deal proposed in the United States?
- How has artificial intelligence impacted healthcare in the last decade?
- Discuss the role of NATO in the 21st century.
- What were the main issues during the verbal confrontations between the US and North Korea under the Trump administration?
- How did the Fukushima nuclear disaster affect energy policies worldwide?
- What are the ongoing challenges in managing global plastic pollution?
- Describe the rise of electric vehicles and their impact on the global oil industry.
- What strategies have been effective in combating deforestation in the Amazon rainforest?
- How has the gig economy transformed traditional employment models?
- What were the significant outcomes of the COP26 summit?
- How do cryptocurrency regulations differ around the world?
- Describe the impact of remote work on urban and suburban development.
- What are the challenges and successes of the Mars Rover missions?
- How did the Arab Spring reshape politics in the Middle East?
- What are the main goals of the United Nations Sustainable Development Goals (SDGs)?
- How has online education evolved apart from the COVID-19 pandemic?
- What are the implications of quantum computing for data security?
- How have drones been integrated into commercial and military operations?
- What are the ethical considerations of gene editing technologies?
- How has social media influenced political campaigns in the 21st century?
- What were the causes and consequences of the global financial crisis of 2008?
- Discuss the impact of the MeToo movement on global workplace policies.
- What are the technological advancements in renewable energy in the last five years?
- How did the Venice Flood Barriers help combat rising sea levels?
- What are the major factors driving urban sprawl in major cities worldwide?
- How have international space laws evolved with the advent of private space travel?
- What are the diplomatic challenges faced by countries in the Arctic as ice melts?

C Disclaimer

We used ChatGPT to rephrase some of the sentences in this paper. But we ensured (to the best of our extent) that all the content in the paper was original.