# Survival Analysis for Cancers of the Brain, CNS and Bone using Retrieval Augmented Generation on the SEER Database

Jyothi Vaidyanathan<sup>1</sup>, Shourya Gupta<sup>2</sup>, Justin Lee<sup>1</sup>, Srikanth Prabhu<sup>2</sup>, Saptarshi Sengupta<sup>1</sup>\*

<sup>1</sup>San Jose State University, San Jose, CA, United States

<sup>2</sup>Manipal Institute of Technology, Manipal, Karnataka, India

{jyothi.vaidyanathan, justin.lee03, saptarshi.sengupta}@sjsu.edu, 13.shourya@gmail.com, srikanth.prabhu@manipal.edu

#### Abstract

Mortality estimation remains a key issue in cancers affecting the Brain, Central Nervous System (CNS), and Bone, among others. The recent integration of LLM-based reasoning into tools that aid cancer prognosis has been particularly encouraging. This prompts us to examine further their stated efficacy and devise workarounds to reduce hallucinations using retrieval augmented generation. We study the clinical, pathological and demographic logs of patients recorded in the National Institutes of Health (NIH) Surveillance, Epidemiology, and End Results (SEER) database and develop an integrated methodology that is user-friendly and responds to n-shot queries with or without context. We first build a set of custom SEER embeddings using DistilBERT, which we use to test tree-based models in answering 'yes/no' type 5-year survivability questions given patient profiles. We extend the limited binary response capability of the prior models by using TabLLM, HyDE-RAG, and Step-Back RAG on the BCNS cancer data and extend them to Bone Cancer data from SEER using GraphRAG, as the attributes are similar. The conversation-friendly models are able to take different context lengths and types into account and provide reasoning about their responses. We successfully show that the extensive patient records in the SEER database can be utilized to develop a powerful conversational agent that is not only able to classify mortality outcomes but also reason about the response by leveraging latent inter-relationships among the unique clinical variables.

## Introduction

Despite advancements in detection and treatment, cancer's complexity challenges early detection, precise prognosis, and effective treatment. Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has the potential to optimize current oncological treatments and can revolutionize the healthcare industry.

In healthcare, input data encompasses diverse modalities, including text, images, videos, and multimodal data. Traditional machine learning approaches do not integrate the full spectrum of structured and unstructured medical data effectively, thereby failing to capture the intricate biological and clinical complexities of cancer. These findings underscore the critical necessity for advancing ML methodologies, leveraging state-of-the-art deep learning and multimodal data fusion techniques.

Large Language Models (LLMs) are an intriguing development in DL, capable of analyzing vast amounts of structured and unstructured data, such as Electronic Health Records (EHRs), genomic data, medical literature and data from Cyber-Physical Healthcare Systems (CPHS). They can provide task-specific results, enabling development of personalized treatment plans for patients, prognosis, and facilitate accurate and informed natural language interactions. Using data from Surveillance, Epidemiology, and End Results (SEER) (National Cancer Institute 1975-2021), we use Generative AI to predict the survivability of cancer patients suffering from brain and CNS cancers. The user interacts with a front-end chat user interface (UI) that ingests prompts and additional context to extract insights from the data. Building on the results obtained, we extend our work to effectively apply Generative AI for bone cancer prognosis.

Furthermore, LLMs facilitate the translation of intricate analytical outputs into comprehensible language. This capability enables healthcare professionals to effectively interpret findings without necessitating advanced technical expertise, thereby enhancing clinical decision-making. LLMs autonomously capture complex attribute relationships, minimizing manual feature engineering. They can retrieve information from reliable external sources when data is insufficient, enhancing response accuracy. As part of this research, we develop LLM models to answer mortality prediction queries in SEER data.

## **Prior Work**

ML is increasingly used to develop predictive models for prognosis and treatment in healthcare. In (Yu et al. 2023), researchers built several models using the SEER data to forecast 5-year survival in non-metastatic cervical cancer patients, highlighting the ability to interpret complex medical data and improve clinical decision-making.

In the study done by Qiao et al., the SEER dataset was used alongside methods like ElasticNet and Multilayer Perceptron to predict distant metastases of thyroid cancer.(Qiao et al. 2023) A comparative analysis was made, demonstrating that the Random Forest model provided the most accurate predictions. In (Nobin, Rahman, and Alam 2022), the

<sup>\*</sup>Corresponding author. Email: saptarshi.sengupta@sjsu.edu Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

SEER dataset is used by 10 traditional machine learning classifiers to predict the survivability of patients with tonsil cancer, where random forest classifier showed the highest results with 93.88% accuracy. Prediction of Lung metastases (LM) and three-month prognostic factors in hepatocellular carcinoma (HCC) patients is done by researchers in (Alkhawaldeh et al. 2023), where Random Forest, Artificial neural network and Easy Ensemble classifiers are utilized. The Easy Ensemble and Random Forest models demonstrated the best performance in predicting outcomes. Both studies, however, encountered the challenge of imbalanced data and addressed this issue in different ways.

LLMs transform tabular data tasks through usage of NLP and traditional machine learning. Hegselmann et al. (2023) show that by serializing data into natural language, LLMs outperform deep learning methods in zero- and few-shot classification tasks due to their pre-existing knowledge base. The paper 'TABLET: Learning From Instructions For Tabular Data' (Slack and Singh 2023) shows that LLMs improve zero-shot performance by 44% with Flan-T5 and 13% with ChatGPT when given natural language instructions.

Many LLMs have been pre-trained in medical datasets specifically for medical applications, such as Me-LLaMA 13B and Me-LLaMA 70B (Xie et al. 2024). By continuous fine-tuning, they outperform other open-source medical LLMs like GPT-4(OpenAI et al. 2024). Google's Med-PaLM 2(Singhal et al. 2023a) is notable, achieving 86.5% accuracy on medical exam queries and delivering expert-level answers. (Singhal et al. 2023b). Another notable LLM model, OncoGPT(Jia et al. 2024), is fine-tuned for oncology-related predictions, which is able to leverage a dataset of more than 180,000 oncology-related notes.

The evaluation metrics for LLMs differ slightly from those of traditional machine learning models. In (Guinet et al. 2024), researchers have created an automated way of RAG(Lewis et al. 2021) evaluation on specific tasks using synthetic multiple choice questions, using item response theory (IRT). LLMs are difficult to evaluate due to the variety of capabilities and inadequate benchmarks as discussed by (Zheng et al. 2023) The research explores the use of powerful LLMs as automated judges that can be scaled to provide decisions comparable to those made by humans.

# **Our Approach**

We use Large Language Models (LLMs) to explore the survivability prediction problem, subsequently selecting the one that provides the most accurate and contextually relevant responses to the user. To implement the prognosis, the following tools are used - Compute Unified Device Architecture (CUDA) enabled GPUs (T4 and A-100), Google Colab/Jupyter Notebook, Python, PyTorch, LangChain.

Data Acquisition and Pre-processing: The SEER dataset is a comprehensive and continuously updated collection of cancer-focused information from various registries throughout the United States. It includes various details about cancer patients, such as demographic, diagnostic, treatment, and outcome data. To enhance predictive power, we conducted a comprehensive review of existing literature, performed Exploratory Data Analysis (EDA), and applied categorical feature encoding for effective feature selection.

## The features used for the Survivability Task are:

'Site recode ICD-O-3/WHO 2008', 'Patient ID', 'Age recode with single ages and 85+', 'CS version input original (2004-2015)', 'RX Summ–Surg Prim Site (1998+)', 'Year of diagnosis', 'ICD-O-3 Hist/behav', 'CS extension (2004-2015)', 'First malignant primary indicator', 'Grade (thru 2017)', 'CS version input current (2004-2015)', 'Primary Site', 'Laterality', '5 year survivability', 'Sex', 'Race/ethnicity', 'Median household income inflation adj to 2019'

(1) **TabLLM:** TabLLM(Hegselmann et al. 2023) is a model designed for few-shot classification of tabular data. It is used in conjunction with the SEER dataset in order to predict cancer survivability outcomes.



Figure 1: Overview of TabLLM (Hegselmann et al. 2023)

**Data Serialization:** Tabular data is serialized using templates converting rows to sentences. Consider a table with columns 'Degree Program', 'Age', 'Is Part of', and 'Graduation year'. A template converts a row like 20, Math club, 2025 into a sentence: 'A person aged 20 in the Electrical Engineering Program graduates in 2025.' (Figure 1).

**Prompting and Fine-Tuning:** The tabular data for prompting the LLMs has been successfully serialized. The T0-11b model is employed to approach few-shot classification. It uses 4 labeled examples to fine-tune the model for the specific task: "Will this patient survive the next 5 years? Yes or no?". Future work will explore additional shots to further enhance the analysis.

Design, Modeling and Module **Specifications:** TabLLM's architecture is designed to enhance fewshot learning and classification on tabular datasets through several key components. Tabular datasets are serialized for efficient data access and processing. The templates standardize query methods and enhance model interpretability. Scripts automate template creation, data preparation, and compliance with model input criteria. The t-few project folder contains scripts for model training and evaluations. These scripts manage the data flow, LLM interactions, and evaluation metrics to ensure an efficient process from dataset preparation to model evaluation.

(2) Use of Embeddings: This methodology utilizes the capabilities of LLMs to transform structured data into numerical representations appropriate for traditional machine learning models. This approach uses embeddings processed from DistilBERT (Sanh et al. 2020) to train and supervise the various classification models.

The necessary libraries are imported and the Arize client is set up to monitor the model's performance(Arize AI



Figure 2: A Classification Pipeline with Vector Embeddings

2024). The embedding generator is initialized using Arize's EmbeddingGeneratorForTabularFeatures using DistilBERT as the model and setting the tokenizer's maximum length to 512 tokens. The dataset is loaded into a Pandas DataFrame, and its columns are divided into four groups to manage the embedding model's context window size. These groups are then sorted into a dictionary. The embeddings and matching prompts for the training and testing datasets for each group of columns are generated and stored into new columns. Following this, each embedding vector is expanded into separate columns, with each vector dimension becoming its own column, using an explore function. An 80-20 split is utilized for training and testing datasets. Instances of various machine learning models (e.g., XGBClassifier, Random Forest) are created on the training set, and their performance is evaluated using metrics such as Accuracy, Precision, Recall, F1 Score, and AUC (Figure 2).

(3) Retrieval Augmented Generation: Retrievalaugmented generation (RAG) is a technique in NLP that combines retrieval-based and generative models to produce more accurate and contextually relevant responses (Figure 3). It consists of two main components: a retriever and a generator. A retriever searches a large knowledge base for relevant information based on the user's prompt, while a generator uses the retrieved information to generate a contextually relevant response. This combination allows the LLM to incorporate external information in real-time. RAG is especially beneficial for tasks that prioritize factual accuracy and access to up-to-date information. It is a superior approach to TabLLM, as it can retrieve updated data without retraining, unlike TabLLM, and can handle large unstructured datasets.



Figure 3: A General RAG Framework

For brain and CNS cancers, serialized representation of the database was fed as external context to the RAG models.

- Data Preprocessing and Loading: The PDF is loaded and processed using PyPDF2 to extract text which is concatenated into a single string. Patient information is segmented using regular expressions, and each segment corresponds to a distinct patient record.
- Embedding Generation: The *FastEmbedEmbeddings* model from LangChain is used to encode each segmented data as numerical vectors. The embeddings are stored in the Chroma(Chroma AI 2024) Vector Store, enabling efficient similarity searches based on user queries.
- **Retrieval Mechanism:** The Chroma Vector Store enables the retriever to identify similar patient records. For each query, the top 20 relevant records are retrieved, and the query is contextualized using a prompt template to enhance accuracy and comprehension.
- Generative Model Integration: After retrieving records and reformulating queries, the Llama-3(Dubey et al. 2024) model is used to generate coherent answers. Prompt templates are useful for procedurally guiding generation and ensure relevant, structured responses.
- Question Processing and Response Generation: User queries are structured to be compatible with Llama-3 processing, allowing for the retrieval of relevant patient record segments. These embeddings are added to the RAG chain, which generates contextually aligned answers. *RunnablePassthrough* in LangChain integrates retrieval and generation components to streamline the process and deliver the final result.

(3.1) RAG Variations and Optimizations: Two variations of RAG were used: Step-Back, HyDE.

**RAG using Step-Back Prompting (Zheng et al. 2024)**: This technique iteratively refines user queries to enhance response accuracy and contextual alignment. It involves submitting a query, retrieving relevant data, generating a preliminary response, and refining the query through a feedback loop (Figure 4). Users can rephrase queries to extract additional insights to enhance the quality of the queries and the final output. LLama-3 is employed for response generation.



Figure 4: A Step-back prompting example showing an instance of a high-school physics concept (Zheng et al. 2024).

**HyDE-RAG:** Hypothetical Document Embeddings (HyDE-RAG(Gao et al. 2022)) method is used in RAG to improve the quality of document retrieval and response generation (Figure 5). By combining document retrieval

and hypothetical reasoning, HyDE is an advanced and context-aware retrieval system compared to traditional RAG. From a user's initial query, the LLM (e.g. GPT-3) generates a hypothetical response. The generated response replaces the original query as the input for the document retriever. The contextual information is transformed into a numerical vector, facilitating a similarity search within the vector store. This process enables the retriever to identify and retrieve documents that are more contextually relevant to the refined query compared to those obtained from the initial query, thereby enhancing the accuracy and relevance of the retrieved information. The generator (another LLM) processes these retrieved documents and the initial query to generate the final response.



Figure 5: The HyDE model (Gao et al. 2022)

Building on these implementation techniques, we have been working on RAG approaches for Bone cancer prognosis using SEER data, starting with an approach called GraphRAG(Microsoft 2024).

GraphRAG: GraphRAG, a Retrieval-Augmented Generation (RAG) technique developed by Microsoft Research, enhances LLMs in synthesizing complex, proprietary datasets not explicitly trained on. Unlike traditional RAG, which retrieves information based solely on vector similarity, it utilizes LLM-generated knowledge graphs to establish connections between distinct data points through shared attributes. This structured approach enables a deeper contextual understanding, allowing it to extract and link related yet disparate information, resulting in more accurate and insightful responses. Particularly relevant to our research, it facilitates survivability prediction by enabling nonexperts to leverage its capabilities without technical expertise. Compared to other RAG approaches discussed above, GraphRAG enables deeper reasoning for the answers provided and can scale efficiently for larger datasets.



Figure 6: Results of various metrics for TabLLM (4 shots)

**TabLLM:** Figure 6 presents performance metrics, including Micro F1 Score, Macro F1 Score, and Accuracy, across various seeds and a shot count of 4 for the 5-year survivability task using the SEER dataset for TabLLM. Using 4 shots for TabLLM is motivated by the scarcity of labeled data in medical datasets. It serves as a baseline for understanding the model's data requirements, while also ensuring computational efficiency and enabling rapid experimentation. The model performed consistently on different seeds (seed 42 produced the most promising results), however, the baseline achieved with TabLLM was not optimal. It was observed that increasing the number of shots did not lead to a significant improvement in performance.

**DistilBERT Embeddings and Tree-based Models:** The DistilBERT language model was used to construct embeddings from the SEER dataset and evaluated using performance metrics, including Accuracy, Precision, Recall, F1 Score, and ROC AUC. These embeddings are fed into various classifier models such as XGBClassifier, Decision Tree, Random Forest, and Gradient Boosting.



Figure 7: Survivability Task Comparative Analysis

Significance of the Results Obtained for the Survivability Task: In predicting the 5-year survival rate for brain cancer patients, boosting methods such as Gradient Boosting and XGBClassifier outperform other techniques due to their ensemble nature, which enhances accuracy and reduces overfitting by iteratively correcting the mistakes of earlier trees. These models achieve higher ROC AUC values than others through effective handling of different data distributions, as shown in Figure 7. RandomForest performs well by averaging multiple decision trees, allowing it to balance variance and bias.



Figure 8: Metastasis Task Comparative Analysis

Significance of Results for Bone Metastasis Task: The results for Bone Metastasis highlight significant challenges due to the unbalanced dataset, as shown in Figure 8. The dataset has 3361 rows as 'No Metastasis' and only 6 rows as 'Yes Metastasis'. As a result, most classifiers demonstrate

high accuracy but very low precision and recall. Many classifiers achieve ROC AUC scores close to 0.5, which suggests their performance is equivalent to that of random guessing. However, models like AdaBoost and XGBClassifier show better class discrimination with their ROC AUC values exceeding 0.6. Despite low overall performance, GaussianNB demonstrates the ability to identify the minority class.

**Retrieval Augmented Generation:** Two benchmarks were evaluated with Simple RAG to establish baseline performance on Llama-3. When tested using the context from the PubMedQA dataset, which provides 'yes' or 'no' answers based on biomedical literature, the model achieved 64.29% accuracy when answering questions about cancer. For domain-specific long-form cancer-related questions, without external context, the model achieved responses with an average cosine similarity of over 60%. These benchmarks indicate promising results, but suggest the need for further refinement when handling detailed queries.

**Comparative Analysis of the RAG Variants:** Step-Back Prompting RAG achieves the highest accuracy at 0.70, excelling in overall correct classifications (Figure 9). Vanilla RAG has superior recall at 0.81 and the highest F1 score at 0.58. However, all methods struggle with low precision.



Figure 9: Comparative Analysis of RAG Variants



Figure 10: Comparative Analysis of Different Techniques

**Key Takeaways:** The results in Figure 10 indicate that XGBClassifier (Embeddings Approach) outperforms other models with high accuracy (0.8505), precision (0.7407), recall (0.6897), F1 score(0.7143), and ROC AUC(0.9177). The bar chart highlights XGBClassifier's superior performance. However, step-back prompting has broader applications due to its ability to handle diverse query encodings in prompts and interact directly with the user. SimpleRAG also provides the best precision-recall balance.

**Preliminary results on Bone Cancer using GraphRAG and Pinecone:** We implemented GraphRAG on the complete data for bone cancer logs in the SEER database. It provides a robust solution with detailed responses, especially when cost is not a limiting factor. We experiment with both global and local search options provided by GraphRAG using the prompt 'What are the common primary sites for a white male?'. Figures 11-12 demonstrate that the local search method deliver more precise results for specific, narrowly-defined queries, making them suitable for use cases requiring tailored answers. We also implemented a chat application with a vector database Pinecone (Pinecone 2025). In Figure 13, significant cases can be attributed to patients in the mentioned age group.



Figure 11: Local Search Prompt/Response with GraphRAG





Figure 12: Global Search Prompt/Response with GraphRAG

# **Conclusion and Future Work**

For survival analysis in the Brain and CNS cancer patients, the ensemble XGBClassifier using custom SEER embeddings outperforms other models. However, step-back prompting allows for interaction with the user and enhances query encoding. For bone cancer, preliminary work indicates GraphRAG excels in delivering detailed, context-rich responses, enabling deeper reasoning and precise information retrieval. These experiments demonstrate the feasibility of posing the cancer survival prediction problem as an LLM-driven natural-language understanding and response generation task, although rigorous benchmarking needs to be done to verify the usefulness, alignment and reliability of the said models. Generative models allow for detailed responses when working with large datasets and also enable integration with external sources to retrieve relevant answers if they are not available within the data. The next step is to evaluate responses using established metrics and fine-tune the models. A comprehensive evaluation of the responses may be done using the 'LLM as a Judge' (Zheng et al. 2023) framework and multiple LLM models can be employed to ensure robustness in the assessments. Future work will aim at expanding the use of LLM-based CPHS to incorporate additional health indicators outside of what is available in the SEER data for cancers of the Brain, CNS and Bone.

# Acknowledgements

The research reported in this publication was supported by the Division of Research and Innovation at San José State University under Award Number 23-UGA-08-044. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of San José State University.

#### References

Alkhawaldeh, I. M.; Altarawneh, G.; Al-Jafari, M.; Abdelgalil, M. S.; Tarawneh, A. S.; and Hassanat, A. 2023. A Machine Learning Approach for Predicting Lung Metastases and Three-Month Prognostic Factors in Hepatocellular Carcinoma Patients Using SEER Data. In 2023 IEEE Symposium on Computers and Communications (ISCC), 1–5.

Arize AI. 2024. Arize Documentation. Available: https://docs.arize.com/arize. Accessed: Jul. 17, 2024.

Chroma AI. 2024. Chroma Documentation: Getting Started. Available: https://docs.trychroma.com/docs/ overview/getting-started. Accessed: Feb. 07, 2025.

Dubey; et al. 2024. The llama 3 herd of models. *arXiv* preprint arXiv:2407.21783.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv:2212.10496.

Guinet, G.; Omidvar-Tehrani, B.; Deoras, A.; and Callot, L. 2024. Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. arXiv:2405.13622.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, 5549–5581. PMLR.

Jia, F.; Liu, X.; Deng, L.; Gu, J.; Pu, C.; Bai, T.; Huang, M.; Lu, Y.; and Liu, K. 2024. OncoGPT: A Medical

Conversational Model Tailored with Oncology Domain Expertise on a Large Language Model Meta-AI (LLaMA). arXiv:2402.16810.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Microsoft. 2024. GraphRAG Documentation. Available: https://microsoft.github.io/graphrag/. Accessed: Feb. 7, 2025.

National Cancer Institute. 1975-2021. Cancer statistics. https://seer.cancer.gov/data/. Accessed: 2025-02-07.

Nobin, R. H.; Rahman, M.; and Alam, M. J. 2022. Survivability Prediction for Patients with Tonsil Cancer Utilizing Machine Learning Algorithms. In 2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), 210–215.

OpenAI; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pinecone. 2025. Pinecone. https://www.pinecone.io. Accessed: 2025-02-07.

Qiao, L.; Li, H.; Wang, Z.; Sun, H.; Feng, G.; and Yin, D. 2023. Machine learning based on SEER database to predict distant metastasis of thyroid cancer. *Endocrine*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; and et al., D. N. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Slack, D.; and Singh, S. 2023. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*.

Xie, Q.; Chen, Q.; Chen, A.; Peng, C.; Hu, Y.; Lin, F.; Peng, X.; Huang, J.; Zhang, J.; and et al.., V. K. 2024. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*.

Yu, W.; Lu, Y.; Shou, H.; Xu, H.; Shi, L.; Geng, X.; and Song, T. 2023. A 5-year survival status prognosis of nonmetastatic cervical cancer patients through machine learning algorithms. *Cancer Medicine*.

Zheng, H. S.; Mishra, S.; Chen, X.; Cheng, H.-T.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. arXiv:2310.06117.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.