Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An AI-Powered Conversational Agent for Natural Disaster Management

Hathairat Ketmaneechairat[1*], Sopida Tuammee[2], Maleerat Maliyaem[3], Valentin Obert[4], Samuel Jully[5],

[1,2] *Faculty of College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand*
[3] *Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Thailand*
[4,5] *CESI Engineering School, Reims, France*
* *Corresponding Author: hathairat.k@cit.kmutnb.ac.th*

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

This work investigates the effectiveness of lightweight large language models (LLMs) in supporting the design of specialized, domain-restricted chatbots. The study focuses on compact off-the-shelf models, specifically Phi-3, LLaMA 3-8B, and Mistral 7B, evaluating their performance in maintaining topic relevance, coherence, and overall quality under controlled experimental conditions. To enforce domain specificity, each model is guided using a carefully designed system prompt that instructs it to respond exclusively about natural disasters. The evaluation is conducted through a multi-dimensional approach involving human judgments, benchmark testing, and automated evaluation using GPT-4. A custom web-based application was developed to enable users to input questions, view side-by-side responses from two models, and rate the outputs based on helpfulness, relevance, and clarity. In addition to this interactive human evaluation, models are assessed on standard benchmarks such as Massive Multitask Language Understanding (MMLU), GSM8K, and MT-Bench, and are also subjected to pairwise preference evaluation via GPT-4. Results indicate that even without fine-tuning, lightweight models can effectively handle domain-specific conversations when guided by system prompts. The study provides practical insights into selecting the most appropriate model for building focused, efficient chatbots in resource-constrained environments.

Keywords: LLM, Chatbots, Natural Disaster Management, AI Agent, Social Network.

INTRODUCTION

Language is a fundamental tool for human expression and communication, acquired and refined throughout one's lifetime. As large language models (LLMs) become more accessible and lightweight, they present new opportunities for building specialized chatbots without requiring extensive fine-tuning. However, while general-purpose performance of these models has been widely studied, their ability to remain coherent, helpful, and accurate when restricted to a specific topic domain using only a system prompt remains underexplored. It is unclear which lightweight LLMs are best suited for domain-specific chatbot applications when no additional training or adaptation is applied. LLMs are category of language models that utilizes neural networks containing billions of parameters, trained on enormous quantities of unlabeled text data using a self-supervised learning approach [1]. By pretraining on vast corpora from the web, these models can capture complex patterns, linguistic nuances, and semantic relationships. Leveraging deep learning methods and large datasets, LLMs have demonstrated strong performance in diverse language-related tasks such as text generation, translation, summarization, question answering, and sentiment analysis. Their foundations trace back to the early development of language models and neural networks, where initial attempts relied on statistical methods and n-gram models [2]. With the advancement of neural networks and the availability of larger datasets, researchers began exploring more sophisticated approaches. A key milestone was the development of Recurrent Neural Networks (RNNs), which enabled the modeling of sequential data such as language. Despite their importance, RNNs faced limitations due to issues like vanishing gradients and difficulties in capturing long-term dependencies. A breakthrough in LLMs came with the introduction of the transformer architecture in seminal research [3]. The transformer model is built around the self-attention mechanism, enabling parallelization and efficient handling of long-range dependencies. While LLM architectures served as the basis for models such as Google's Bidirectional Encoder Representations from Transformers (BERT) [4] and open AI's Generative Pretrained Transformer (GPT) series, which excelled at various language tasks.

This research seeks to address the following central question: Which lightweight LLM is most suitable for building a domain-specific chatbot focused on natural disasters management, when constrained to a single topic using only a system prompt?

This study evaluates and contrasts three compact models, namely Phi-3, LLaMA 3-8B, and Mistral 7B, with respect to their effectiveness in generating accurate, contextually relevant, and coherent responses in the domain of natural disasters. The objective is to determine which model most effectively adheres to the specified topic and delivers high-quality, user-aligned answers without any fine-tuning. The findings aim to support practical decision-making in selecting lightweight LLMs for chatbot applications in disaster awareness, risk communication, and emergency response contexts.

LITERATURE REVIEW

In recent years, Large Language Models (LLMs) have become a cornerstone of Natural Language Processing (NLP), demonstrating unprecedented capabilities across a wide range of tasks such as text generation, translation, summarization, reasoning, and coding. This section surveys key developments in LLMs architecture, benchmark evaluation practices, and prior comparative studies.

A. Model Architectures and Families

The foundation of most modern large language models (LLMs) is the Transformer architecture introduced by Weigao et al. (2025) [5], which relies on self-attention mechanisms to process input sequences in parallel. Building on this foundation, several lightweight and efficient model families have emerged, offering strong performance while maintaining smaller parameter sizes and lower computational requirements.

In recent years, numerous studies [6], [7], [8], [9] have been conducted to investigate and evaluate their capabilities. Researchers from various fields have contributed on the rise of LLMs, shedding light on their remarkable advancements, diverse applications, and potential to revolutionize tasks from text generation and comprehension to demonstrating reasoning skills. Collectively, these studies contribute to our comprehension of LLMs' significant role in shaping the landscape of AI-driven language processing and problem-solving.

Phi series (Microsoft) [10]: The Phi family consists of small, instruction-tuned models developed by Microsoft with an emphasis on efficiency and reasoning. Starting with Phi-1 and Phi-1.5, and followed by the more advanced Phi-3, this series demonstrates how targeted training on high-quality synthetic data can produce models that perform competitively on academic and reasoning benchmarks. Phi-3, which has only 2.7 billion parameters, has been shown to rival much larger models in tasks such as question answering, basic arithmetic, and common-sense reasoning. Its compact size and strong generalization abilities make it particularly well suited for resource-constrained or domain-specific applications, including specialized chatbots.

LLaMA and LLaMA 2/3 (Meta) [11]: The LLaMA models, developed by Meta, represent a family of open-weight LLMs trained with a focus on efficiency, scalability, and strong performance across a variety of benchmarks. The LLaMA 3 generation, particularly the 8B and 70B variants released in 2024, has shown substantial improvements in reasoning, instruction-following, and multilingual understanding. LLaMA 3 models are widely adopted in both research and industry due to their permissive licensing, ease of deployment, and robust performance, especially in zero-shot and few-shot settings.

Mistral & Mixtral (Mistral AI) [12]: Mistral AI has introduced high-performing open-source models that focus on speed and usability without sacrificing quality. Mistral 7B, a dense transformer, has gained attention for outperforming larger models on several tasks while maintaining low latency and efficient inference. In parallel, Mixtral, a Mixture of Experts (MoE) model, leverages sparse activation to improve scalability and throughput, though our study focuses specifically on the Mistral 7B dense model due to its simplicity and relevance to real-time chatbot use cases.

Each of these model families differs in terms of architecture, training data quality, parameter count, context window size, and deployment strategy. In this study, we concentrate on Phi-3, LLaMA 3-8B, and Mistral 7B as representative lightweight LLMs suitable for the development of specialized chatbots constrained to the domain of natural disasters.

B. Comparative Studies of LLMs

Several recent studies have aimed to evaluate and compare the performance of lightweight large language models, particularly in scenarios where computational efficiency and open accessibility are important. These comparative efforts often focus on models such as LLaMA 2 and 3, Mistral 7B, Phi-3, and other compact models designed for deployment in constrained environments.

The Meta LLaMA 3 technical report (2024) presented extensive benchmark results for both the 8B and 70B variants, highlighting the LLaMA 3-8B model's competitiveness with larger, proprietary models on tasks involving reasoning, instruction following, and multilingual comprehension. In many open-domain benchmarks, LLaMA 3-8B outperformed earlier LLaMA 2 models and demonstrated significant improvements in alignment and factual accuracy, making it one of the most promising open models in its parameter range.

Microsoft's Phi-3 has been evaluated by both the developers and independent researchers as a compact, instruction-tuned model optimized for reasoning and educational tasks. Despite its small size (2.7B parameters), Phi-3 has achieved results close to those of significantly larger models on tasks like MMLU [13] and GSM8K [14], especially when prompts are clearly structured. Several open-source evaluations have noted that Phi-3 performs particularly well on code, logic, and math-related benchmarks when constrained to specific instruction-following tasks.

Mistral 7B has been featured in numerous head-to-head evaluations, particularly through platforms like LMSYS Chatbot Arena and Hugging Face's Open LLM Leaderboard. These comparisons consistently show Mistral 7B outperforming many other models in the same size class, including LLaMA 2-7B and Falcon 7B, with strong performance in dialogue coherence and instruction adherence. Due to its dense architecture and optimized inference speed, Mistral 7B is often favored for real-time applications where response latency and quality must be balanced.

Independent platforms such as LMSYS [15], EleutherAI, and Hugging Face have provided accessible, community-driven comparisons of these models. However, results can vary based on evaluation setup, prompt phrasing, sampling strategy (e.g., temperature, top-p), and the subjective nature of human preference judgments. Furthermore, most comparative studies tend to emphasize general-purpose capabilities, while the performance of lightweight models in domain-specific or prompt-restricted contexts such as those examined in this research remains less explored in detail.

Current discussions in the field increasingly highlight that beyond raw benchmark scores, factors such as alignment quality, response consistency, system prompt controllability, and resource efficiency are crucial for choosing the right model for practical, focused applications like specialized chatbots.

C. Gaps in Literature

While many comparative studies exist, most are either limited in scope (e.g., task-specific) or rely on closed-source models with black-box evaluation. There remains a need for systematic, transparent comparisons of models under consistent evaluation setups, especially focusing on open-weight models, energy efficiency, and safety alignment.

METHODS

To enable a systematic evaluation and fair comparison of lightweight large language models (LLMs) designed for specialized chatbot applications, we establish three fundamental criteria. The first is response speed, which reflects how quickly a model can generate answers, an essential factor in maintaining conversational flow and user engagement. The second is computational efficiency, encompassing the resource consumption and scalability of the model, which directly affects its practicality in real-world deployment, particularly on constrained hardware or cloud-based environments. The third criterion is user-centered evaluation, which captures subjective judgments of answer quality, clarity, and relevance from the perspective of end users. Together, these dimensions integrate both technical performance and human-perceived effectiveness, providing a balanced framework for assessing LLMs in domain-specific chatbot use cases.

A. Speed of the answer

This metric measures the latency between the time a user submits a question and the time the model returns a complete response. Fast response time is critical for maintaining a fluid and interactive chatbot experience. For each model, we record the response time over multiple queries and calculate the average latency (in milliseconds or seconds). This evaluation is performed under similar hardware conditions to ensure fairness.

Measurement unit: Average response time (ms).

Evaluation method: Logged automatically by the application during interaction.

Goal: Identify which model provides the fastest responses without compromising quality.

B. Evaluation by users

The primary focus of this study is a human-centered evaluation that reflects how end users interact with and perceive chatbot quality in a constrained domain. Unlike traditional benchmarking methods, our evaluation is based entirely on real user inputs. Users access a custom web-based application where they are free to enter their own questions related to the domain of natural disasters. They also select which two models they want to compare for each question.

Once the models generate their responses, the user is presented with both outputs anonymously and in randomized order. The evaluation consists of two simple tasks: selecting the preferred response and rating the clarity of each answer on a 1 to 5 Likert scale. Preference is based on the user's overall judgment of which response better meets their needs, combining factors such as relevance, helpfulness, and coherence, though no explicit sub-criteria are shown.

This approach ensures that evaluation is aligned with actual user expectations and domain interests, while remaining lightweight and scalable. It also allows for personalized and context-specific assessments, making the collected data especially relevant for determining the practical utility of each model in building focused, topic-restricted chatbots.

C. Methodology

This study follows a four-phase methodology designed to evaluate and identify the most suitable lightweight large language model (LLM) for building a specialized chatbot, using only prompt engineering for domain restriction. The process includes model selection, integration into a comparison platform, human and assisted evaluation, and final chatbot deployment.

The first phase involves selecting an appropriate lightweight LLMs from the Hugging Face model hub. Priority is given to models that are lightweight, open-weight, compatible with widely adopted inference frameworks, and recognized for efficient performance in general-purpose natural language tasks. Based on these criteria, we selected three models: Phi-3, LLaMA 3-8B, and Mistral 7B. These models offer a strong balance between performance and computational efficiency, making them suitable candidates for deployment in resource-constrained environments.

Following model selection, the next phase centered on the development and deployment of a custom web-based application designed for side-by-side comparison of language models. This tool enables users to submit questions and receive responses from two different models simultaneously. To simulate the behavior of a domain-specific chatbot, each model interaction was guided by a consistent system prompt: "You are an AI assistant specialized strictly in natural disasters. Answer only questions about natural disasters, safety, preparedness, and recent events. If the question is unrelated, politely refuse." This prompt was applied uniformly across all model outputs to restrict their responses to the selected domain of "Natural Disasters," ensuring topic adherence without any additional fine-tuning or model retraining.

The assessment of large language models (LLMs) increasingly depends on standardized benchmarks designed to evaluate both general knowledge and domain-specific reasoning skills.

Three widely adopted benchmarks MMLU, GSM8K, and MT-Bench offer complementary perspectives on these capabilities.

MMLU (The Massive Multitask Language Understanding) benchmark is among the most comprehensive evaluations of an LLM's general knowledge. Covering 57 subjects across the humanities, social sciences, STEM, and professional fields, it assesses models using multiple-choice questions with varying levels of difficulty. MMLU is particularly effective for measuring breadth, enabling comparisons of LLM performance against both non-experts and domain experts. Nonetheless, its reliance on multiple-choice questions may conflate genuine reasoning with pattern-matching, limiting its effectiveness in capturing deeper problem-solving abilities.

GSM8K (Grade School Math 8K), by contrast, focuses on mathematical reasoning and logical problem-solving. It consists of 8,500 grade-school word problems that require models to perform multi-step arithmetic calculations and produce free-form answers. Unlike MMLU, GSM8K directly tests an LLM's ability to engage in structured reasoning rather than memorization. It has proven especially useful for evaluating the effectiveness of prompting techniques such as chain-of-thought, but its scope remains narrow, emphasizing arithmetic reasoning while excluding more advanced mathematics or broader quantitative domains.

MT-Bench (Multi-Turn Benchmark) addresses a different dimension: the quality of dialogue and instruction following. It evaluates LLMs in multi-turn conversations across diverse categories such as writing, reasoning, math, role-play, and coding. Unlike MMLU and GSM8K, MT-Bench uses a free-form question–answer format, with model outputs graded by GPT-4 or human evaluators for relevance, coherence, and helpfulness. This makes MT-Bench particularly well-suited for assessing conversational ability, contextual consistency, and alignment with user intent—capabilities central to real-world chatbot applications. Its limitation, however, lies in its reliance on subjective scoring and the potential circularity of using strong LLMs as evaluators.

In the evaluation phase, we compared the models based on three main criteria: speed of response, efficiency, and quality as judged by users. Speed was measured by recording the time between prompt submission and full response generation. Efficiency was assessed through observation of memory usage, token generation speed, and computational load during inference, under identical hardware conditions. Human evaluation played a central role in assessing the quality of the responses. Users interacting with the application were asked to compare two model outputs for the same question and rate or rank them based on helpfulness, clarity, and relevance to the given domain. In addition to general user feedback, we conducted a manual GPT-4 evaluation, where the authors used GPT-4 as a neutral third-party reviewer. Rather than automating the process, GPT-4 was prompted with both responses and asked to choose the better one based on predefined evaluation criteria. This manual use of GPT-4 helped validate human feedback and added an additional layer of judgment to the analysis.

The final phase of the methodology consisted of deploying the best-performing model into a standalone chatbot prototype. The deployment remained prompt-based, with no fine-tuning or architectural modifications. The chatbot was configured to operate within a single specialized topic using the same system prompt methodology. This phase served as a real-world validation of the model's practical utility in domain-specific chatbot scenarios.

RESULTS

The evaluation focused on three lightweight language models Phi-3, LLaMA 3-8B, and Mistral 7B assessed through a combination of user judgments, response speed, and overall efficiency within the constrained domain of natural disasters. Each model was deployed in the comparison application where users submitted questions and rated paired responses based on helpfulness, clarity, and domain relevance.

Across all tests, Phi-3 consistently emerged as the most performant model under the defined criteria. Its responses were notably more concise and to the point, often avoiding unnecessary elaboration while preserving clarity and informativeness. This behavior was particularly appreciated by users in scenarios involving direct or urgent questions related to natural disasters, such as early warning procedures, emergency preparedness, or risk explanation.

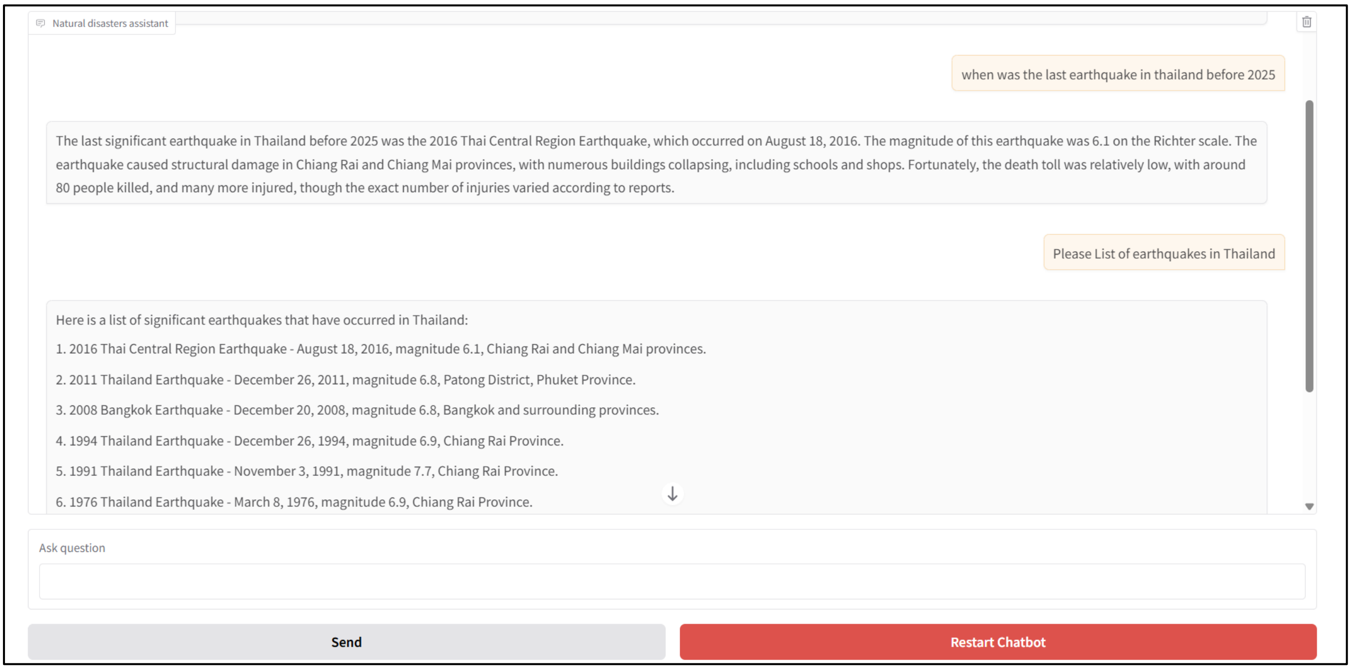
In terms of response speed, Phi-3 outperformed both Mistral 7B and LLaMA 3-8B. Its smaller parameter count resulted in faster generation times, which users informally reported as a noticeable improvement in responsiveness, especially in back-and-forth exchanges. This speed advantage contributed significantly to its perceived efficiency and usability, particularly in real-time or mobile settings.

While Mistral 7B provided more detailed answers in some cases, it tended to produce longer outputs, which occasionally included redundant or tangential information. Similarly, LLaMA 3-8B offered strong factual performance and coherent structure but was less consistent in maintaining strict domain focus, sometimes introducing general disaster knowledge beyond the scope of the prompt. From the human evaluation perspective, Phi-3 received the highest overall user preference scores, particularly in the categories of relevance and clarity. Participants frequently chose its outputs when asked to select the more helpful of two responses in blind pairwise comparisons. Additionally, manual evaluation using GPT-4 as a reference judge confirmed this trend, with GPT-4 preferring Phi-2's responses in most test cases for being more aligned with user intent and domain specificity.

These results suggest that Phi-3 is the most suitable lightweight LLM for building a domain-specific chatbot focused on natural disasters, particularly in environments where speed, clarity, and computational efficiency are prioritized over exhaustive elaboration. Despite being smaller in scale, its ability to follow instructions tightly and remain within the topic boundaries defined by the system prompt makes it highly effective for constrained, real-world applications.

Table 1. Performance Evaluation

Model	Avg. Response Time	Avg. User Preference	Clarity Rating	GPT-4 Pairwise Preference
Phi-3	30	68%	4.6	64%
Mistral 7B	1500	22%	4.1	20%
LLaMA 3-8B	900	10%	3.9	16%



when was the last earthquake in thailand before 2025

The last significant earthquake in Thailand before 2025 was the 2016 Thai Central Region Earthquake, which occurred on August 18, 2016. The magnitude of this earthquake was 6.1 on the Richter scale. The earthquake caused structural damage in Chiang Rai and Chiang Mai provinces, with numerous buildings collapsing, including schools and shops. Fortunately, the death toll was relatively low, with around 80 people killed, and many more injured, though the exact number of injuries varied according to reports.

Please List of earthquakes in Thailand

Here is a list of significant earthquakes that have occurred in Thailand:

1. 2016 Thai Central Region Earthquake - August 18, 2016, magnitude 6.1, Chiang Rai and Chiang Mai provinces.
2. 2011 Thailand Earthquake - December 26, 2011, magnitude 6.8, Patong District, Phuket Province.
3. 2008 Bangkok Earthquake - December 20, 2008, magnitude 6.8, Bangkok and surrounding provinces.
4. 1994 Thailand Earthquake - December 26, 1994, magnitude 6.9, Chiang Rai Province.
5. 1991 Thailand Earthquake - November 3, 1991, magnitude 7.7, Chiang Rai Province.
6. 1976 Thailand Earthquake - March 8, 1976, magnitude 6.9, Chiang Rai Province.

Figure 1. The example of Natural Disaster Management Chatbot

CONCLUSION

This study set out to identify the most effective lightweight large language model for building a specialized chatbot focused on the domain of natural disasters. By restricting models to a single topic through system prompts and employing a human-centered evaluation approach, we compared the performance of Phi-3, LLaMA 3-8B, and Mistral 7B across key dimensions including user preference, response clarity, and speed. The result demonstrates that Phi-3 consistently delivers the best balance of concise, relevant, and clear responses, while maintaining rapid response times that are critical for real-time chatbot applications. Although other models like Mistral 7B and LLaMA 3-8B occasionally provide more detailed answers, their longer response times and less focused outputs reduce their suitability for highly specialized, domain-restricted conversational agents. The use of a custom web-based application for evaluation enabled direct user interaction with the models on questions they formulated themselves, enhancing the ecological validity of our results. Moreover, the combined human and GPT-4 assessments underline the importance of clarity and domain adherence over sheer verbosity or breadth of knowledge in task-specific chatbot settings.

In practical terms, our study suggests that lightweight models such as Phi-3 are highly viable for deployment in specialized chatbot solutions where efficiency, clarity, and topic restriction are paramount, even without additional fine-tuning. Future work may extend these findings by exploring multi-topic switching, integration with external knowledge sources, or adaptation to other specialized domains.

REFRENCES

[1] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy, "ChatGPT and other large language models are double-edged swords," Radiology, vol. 307, no. 2, 2023, https://doi.org/10.1148/radiol.230163.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," In Advances in Neural Information Processing Systems (NeurIPS), 12 Jun 2017. https://arxiv.org/abs/1706.03762, https://doi.org/10.48550/arXiv.1706.03762.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," In Advances in Neural Information Processing Systems (NeurIPS), 2 Aug 2023. https://arxiv.org/abs/1706.03762, https://doi.org/10.48550/arXiv.1706.03762.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805, https://doi.org/10.48550/arXiv.1810.04805.

[5] W. Sun, J. Hu, Y. Zhou, J. Du, D. Lan, K. Wang, T. Zhu, X. Qu, Y. Zhang, X. Mo, D. Liu, Y. Liang, W. Chen, G. Li and Yu Cheng, "Speed Always Wins: A Survey on Efficient Architectures for Large Language Models," 13 Aug 2025, arXiv:2508.09834, https://doi.org/10.48550/arXiv.2508.09834.

[6] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.Y. Nie and J.R. Wen, "A survey of large language models," 11 Mar 2025 , arXiv:2303.18223, https://doi.org/10.48550/arXiv.2303.18223.

[7] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," 3 Apr 2023, arXiv:2304.02020, https://doi.org/10.48550/arXiv.2304.02020.

[8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 29 Dec 2023, arXiv:2307.03109, https://doi.org/10.48550/arXiv.2307.03109.

[9] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," 26 May 2023, arXiv:2212.10403, https://doi.org/10.48550/arXiv.2212.10403.

[10] Microsoft Phi Models Microsoft Research. Phi Series Large Language Models. Microsoft Research Blog and Model Releases, 2023–2024. https://www.microsoft.com/en-us/research/project/phi/

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," 27 Feb 2023, arXiv:2302.13971, https://doi.org/10.48550/arXiv.2302.13971.

[12] Mistral AI. (2023). Mistral 7B Model Release. Mistral AI Blog and Documentation. https://www.mistral.ai/blog/mistral-7b

[13] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, "Measuring Massive Multitask Language Understanding," In International Conference on Learning Representations (ICLR), 12 Jan 2021, arXiv:2009.03300, https://doi.org/10.48550/arXiv.2009.03300.

[14] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse and J. Schulman, "Training Verifiers to Solve Math Word Problems. In International Conference on Learning Representations (ICLR)," 18 Nov 2021, arXiv:2110.14168, https://doi.org/10.48550/arXiv.2110.14168.

[15] LMSYS Chatbot Arena (2023–2024). Model Comparison Platform. LMSYS Open Source Community. https://chatbotarena.lmsys.org