Large Language Models: Literature Review

June 18, 2024

Abstract

This literature review explores the advancements in large language models (LLMs), with a focus on text reasoning tasks, multilingual model performance, and open-source LLM code datasets. The review critically analyzes key contributions, methodologies, and limitations in these areas. It discusses the implications of the research for the development and deployment of LLMs, highlighting the need for enhanced efficiency, reliability, and inclusivity in natural language processing.

1 Introduction

Large language models (LLMs) have transformed the field of natural language processing (NLP) by demonstrating significant capabilities in various tasks. This literature review aims to provide a comprehensive overview of the advancements in LLMs, focusing on their performance in text reasoning tasks, multilingual modeling, and their availability as open-source code datasets. The review also discusses the challenges and implications of these developments.

2 Background

The evolution of LLMs has been marked by the increasing complexity and scale of models, resulting in significant improvements in NLP tasks. However, these advancements come with challenges such as high computational costs and unpredictable outputs. Understanding these models' development and performance across different contexts is crucial for optimizing their use in diverse applications.

3 Literature Review

3.1 LLMs in Text Reasoning Tasks

Large Language Models (LLMs) have demonstrated noteworthy capabilities in text reasoning tasks, such as spatial reasoning, as evidenced by Claude 3's performance on tasks it likely never encountered during training [Greatrix et al., 2024]. However, their application in safety-critical scenarios remains problematic due to unpredictable hallucinations and misinformation [Qiu and Miikkulainen, 2024]. New frameworks like Semantic Density aim to mitigate these issues by extracting uncertainty information but are limited to specific tasks [Qiu and Miikkulainen, 2024].

LLMs like ChatGPT have shown impressive results in various Natural Language Processing (NLP) tasks, but systematic investigations into their potential and limitations are still ongoing. Recent studies provide comprehensive overviews and propose taxonomies for better application [Qin et al., 2024a]. Despite this, LLMs often struggle with tasks requiring extensive real-world knowledge and handling long-tail entities, highlighting the need for supplemental non-parametric knowledge [Huang et al., 2024].

Evaluations of LLMs on commonsense reasoning, mathematical reasoning, and natural language generation (NLG) reveal their potential yet highlight their constraints in reliably generating factual content [Ni and Li, 2024]. To address these limitations, recent works propose employing retrieval-augmented methods or novel frameworks like Knowledge Selection LLMs (KS-LLM) to select relevant knowledge snippets for answering questions accurately [Zheng et al., 2024].

LLMs also excel in complex text generation tasks, such as summarization and dialogue generation, but often suffer from hallucinations or non-factual outputs [Jacoby et al., 2024, Shi et al., 2024]. Methods like Learnable Intervention for Truthfulness Optimization (LITO) aim to improve accuracy by guiding LLM responses towards factual consistency [Bayat et al., 2024].

To further enhance their reasoning capabilities, approaches like chain-of-thought (CoT) and tree-of-thought (ToT) methods have been explored, where problems are decomposed into sequences of reasoning steps [Kang et al., 2024]. Additionally, leveraging external reasoning frameworks, such as Lean theorem proving, has shown promise in addressing logical inconsistencies and improving complex reasoning outcomes [Jiang et al., 2024].

Lastly, ensuring the robustness of LLMs in adversarial settings continues to be a research focus, with innovative fine-tuning techniques and new metrics being introduced to evaluate and enhance their reliability [Yang et al., 2024, Siegel et al., 2024]. The development of benchmarks like CFLUE underline the ongoing efforts to keep pace with the rapid advancements in LLMs [Zhu et al., 2024].

3.2 Multilingual Model Performance

Multilingual performance of Large Language Models (LLMs) has seen significant enhancements through various methodologies. Instruction-tuning with question translation data without annotated answers has proven effective across multiple languages, even those not included in the instruction-tuning phase [Zhang et al., 2024a]. Open-source LLMs, trained on high-quality datasets comprising 70k prompt-response pairs across 74 languages, have outperformed prior models in multilingual benchmarks, highlighting the benefits of multilingual data for specific languages such as Japanese [Devine, 2024].

Challenges exist in training LLMs in non-English languages due to the scarcity of large-scale corpora and needed computing resources. Innovative approaches like ChatFlow for Chinese-English cross-language transfer leverage mixed corpora to facilitate knowledge transfer and optimize language-specific models [Li et al., 2024b]. Specific languages, such as Romanian, have seen specialized foundational chat LLMs developed to improve performance in non-English contexts [Masala et al., 2024].

Low-resource languages, including indigenous groups like the Sami and underrepresented languages such as Turkish, face unique challenges due to data scarcity. Efforts focus on enhancing multilingual LLM performance through fine-tuning and creating instructional datasets tailored to these languages [Paul et al., 2024, Acikgoz et al., 2024]. In particular, techniques like Dynamic Preference Optimization for Italian and federated learning for decentralized data utilization hold promise for advancing multilingual capabilities [Polignano et al., 2024, Sani et al., 2024].

The introduction of robust, open-sourced multilingual LLMs like Tele-FLM, which utilize an efficient pre-training paradigm, further underscores the potential for superior multilingual language modeling [Li et al., 2024a]. Studies also highlight the efficiency of large-scale models like Claude 3 in machine translation tasks, emphasizing the role of synthetic data and knowledge distillation in improving translation quality for low-resource languages [Enis and Hopkins, 2024].

Newly developed models, such as BLOOMZMMS for multilingual speech recognition and CT LLM 2B for Chinese language prioritization, demonstrate the evolving focus on leveraging the vast linguistic data for enhancing multilingual performance [Denisov and Vu, 2024, Du et al., 2024]. To support low-resource language inclusion, initiatives like EthioLLM for Ethiopian languages and the development of tailored synthetic data approaches are critical [Tonja et al., 2024, Doshi et al., 2024].

Overall, the continuous improvement in multilingual LLMs requires addressing data scarcity, enhancing fine-tuning methods, and innovating training strategies to ensure robust performance across diverse languages [Qin et al., 2024b, Chen and Li, 2024]. These endeavors contribute significantly to the inclusivity and capability of LLMs in performing across a wide variety of linguistic contexts.

3.3 Performance of Large Language Models

The rapid advancement of large language models (LLMs) has led to significant improvements in natural language processing while also posing challenges due to high computational and energy demands ([Hillier et al., 2024, Chen et al., 2024, Ji et al., 2024, Moar et al., 2024]). Researchers have explored innovative techniques like byte-level tokenization, pooling mechanisms, weight tying, efficient training strategies, and low-rank compression to reduce parameter counts by 90-95

The performance of LLMs has been shown to excel across a variety of tasks, yet the memorization of training data and the efficient handling of long contexts pose additional challenges ([Schwarzschild et al., 2024, Chavan et al., 2024]). Techniques like model compression, quantization, and pruning have been actively explored to address these challenges, often trading off between accuracy and efficiency ([Dotzel et al., 2024, He and Wu, 2024, Zou et al., 2024]). New models, such as OpenBA V2, MiniCPM, and Hierarchical Context Merging (HOMER), have demonstrated significant compression with minimal performance loss, providing scalable solutions for resource-constrained environments ([Hu et al., 2024, Recasens et al., 2024, Tyukin, 2024]).

In summary, while LLMs have revolutionized natural language processing, their high computational demands necessitate ongoing innovation in efficiency techniques to maximize their practical application across diverse tasks and scenarios ([Qi et al., 2024]).

3.4 Open-Source LLM Code Datasets

Open-ended inquiries with LLMs face issues such as hallucinations and high training costs. Addressing these, techniques include augmenting inputs with knowledge graphs, though challenges remain in extracting relevant information and differing KG characteristics [Lin et al., 2024b]. The potential of LLMs in specialized professional domains is notable, but limited by the access to closed-source APIs and data acquisition difficulties. Initiatives like CourseGPT-zh aim to provide cost-effective LLM deployment tailored to educational contexts by designing specific corpora and optimization frameworks [Qu et al., 2024]. Open-source datasets, like the one presented for the OpenROAD EDA toolchain, are essential to support LLM training in technical fields [Wu et al., 2024]. For mobile deployment, frameworks such as llama-cpp enable on-device operation of LLMs, demonstrating efficient performance on devices like the Galaxy S21 [Fassold, 2024]. Comprehensive evaluation of LLM utility in real-world scenarios is being addressed through datasets like User Reported Scenarios (URS), which collect diverse user cases [Wang et al., 2024b]. Furthermore, the use of LLMs in software engineering, particularly in tasks like SQL query interpretation, highlights the variability in model performance based on token overlap in source code [Zhang et al., 2024b, Haldar and Hockenmaier, 2024]. Evaluating LLM capabilities for processing long documents is crucial, with specific benchmarks like L-Eval and Long-Bench providing structured assessments [Wang et al., 2024a]. Finally, recent approaches to Math Word Problems emphasize the reduction of annotation costs through weak supervision, relying solely on final answer supervision [Lin et al., 2024a].

4 Discussion

The research reviewed highlights the substantial progress in LLMs, particularly in text reasoning and multilingual tasks. However, significant challenges remain, such as mitigating hallucinations, addressing data scarcity, and improving computational efficiency. The effectiveness of LLMs in specialized domains underscores the need for tailored datasets and fine-tuning methods to enhance performance. Additionally, the scalability and deployment costs must be tackled to make LLMs more accessible and practical for broader applications.

5 Conclusion

This literature review synthesizes current research on large language models, focusing on their capabilities and limitations in text reasoning tasks, multilingual performance, and the availability of open-source code datasets. Future research should address the identified challenges, such as improving factual consistency, enhancing multilingual model performance, and optimizing computational efficiency. These efforts will be crucial in advancing the development and deployment of LLMs across diverse applications.

References

Emre Can Acikgoz, Mete Erdogan, and Deniz Yuret. Bridging the bosphorus: Advancing turkish large language models through strategies for low-resource language adaptation and benchmarking, 2024. URL https://arxiv.org/abs/2405.04685.

Farima Fatahi Bayat, Xin Liu, H. V. Jagadish, and Lu Wang. Lito: Learnable intervention for truthfulness optimization, 2024. URL https://arxiv.org/abs/2405.00301.

Arnav Chavan, Nahush Lele, and Deepak Gupta. Surgical feature-space decomposition of llms: Why, when and how?, 2024. URL https://arxiv.org/abs/2405.13039.

Bowen Chen, Namgi Han, and Yusuke Miyao. A multi-perspective analysis of memorization in large language models, 2024. URL https://arxiv.org/abs/2405.11577.

Lung-Chuan Chen and Zong-Ru Li. Bailong: Bilingual transfer learning based on qlora and zip-tie embedding, 2024. URL https://arxiv.org/abs/2404.00862.

Pavel Denisov and Ngoc Thang Vu. Teaching a multilingual large language model to understand multilingual speech via multi-instructional training, 2024. URL https://arxiv.org/abs/2404.10922.

Peter Devine. Tagengo: A multilingual chat dataset, 2024. URL https://arxiv.org/abs/2405.12612.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. Do not worry if you do not have data: Building pretrained language models using translationese, 2024. URL https://arxiv.org/abs/2403.13638.

Jordan Dotzel, Yuzong Chen, Bahaa Kotb, Sushma Prasad, Gang Wu, Sheng Li, Mohamed S. Abdelfattah, and Zhiru Zhang. Learning from students: Applying t-distributions to explore accurate and efficient formats for llms, 2024. URL https://arxiv.org/abs/2405.03103.

Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, Xinchen Luo, Guorui Zhou, Binhang Yuan, Wenhu Chen, Jie Fu, and Ge Zhang. Chinese tiny llm: Pretraining a chinese-centric large language model, 2024. URL https://arxiv.org/abs/2404.04167.

Maxim Enis and Mark Hopkins. From llm to nmt: Advancing low-resource machine translation with claude, 2024. URL https://arxiv.org/abs/2404.13813.

Hannes Fassold. Porting large language models to mobile devices for question answering, 2024. URL https://arxiv.org/abs/2404.15851.

Thomas Greatrix, Roger Whitaker, Liam Turner, and Walter Colombo. Can large language models create new knowledge for spatial reasoning tasks?, 2024. URL https://arxiv.org/abs/2405.14379.

Rajarshi Haldar and Julia Hockenmaier. Analyzing the performance of large language models on code summarization, 2024. URL https://arxiv.org/abs/2404.08018.

Qiaozhi He and Zhihua Wu. Efficient llm inference with kcache, 2024. URL https://arxiv.org/abs/2404.18057.

Dylan Hillier, Leon Guertler, Cheston Tan, Palaash Agrawal, Chen Ruirui, and Bobby Cheng. Super tiny language models, 2024. URL https://arxiv.org/abs/2405.14159.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang

Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z. Pan. Prompting large language models with knowledge graphs for question answering involving long-tail facts, 2024. URL https://arxiv.org/abs/2405.06524.

Derek Jacoby, Tianyi Zhang, Aanchan Mohan, and Yvonne Coady. Human latency conversational turns for spoken avatar systems, 2024. URL https://arxiv.org/abs/2404.16053.

Yixin Ji, Yang Xiang, Juntao Li, Wei Chen, Zhongyi Liu, Kehai Chen, and Min Zhang. Feature-based low-rank compression of large language models via bayesian optimization, 2024. URL https://arxiv.org/abs/2405.10616.

Dongwei Jiang, Marcio Fonseca, and Shay B. Cohen. Leanreasoner: Boosting complex logical reasoning with lean, 2024. URL https://arxiv.org/abs/2403.13312.

Liwei Kang, Zirui Zhao, David Hsu, and Wee Sun Lee. On the empirical complexity of reasoning and planning in llms, 2024. URL https://arxiv.org/abs/2404.11041.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, Shuangyong Song, Yongxiang Li, Zheng Zhang, Bo Zhao, Aixin Sun, Yequan Wang, Zhongjiang He, Zhongyuan Wang, Xuelong Li, and Tiejun Huang. Tele-flm technical report, 2024a. URL https://arxiv.org/abs/2404.16645.

Yudong Li, Yuhao Feng, Wen Zhou, Zhe Zhao, Linlin Shen, Cheng Hou, and Xianxu Hou. Dynamic data sampler for cross-language transfer learning in large language models, 2024b. URL https://arxiv.org/abs/2405.10626.

Qingwen Lin, Boyan Xu, Zhengting Huang, and Ruichu Cai. From large to tiny: Distilling and refining mathematical expertise for math word problems with weakly supervision, 2024a. URL https://arxiv.org/abs/2403.14390.

Yu-Hsiang Lin, Huang-Ting Shieh, Chih-Yu Liu, Kuang-Ting Lee, Hsiao-Cheng Chang, Jing-Lun Yang, and Yu-Sheng Lin. Retrieval-augmented language model for extreme multi-label knowledge graph link prediction, 2024b. URL https://arxiv.org/abs/2405.12656.

Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. Openllm-ro – technical report on open-source romanian llms, 2024. URL https://arxiv.org/abs/2405.07703.

Chakshu Moar, Michael Pellauer, and Hyoukjun Kwon. Characterizing the accuracy - efficiency trade-off of low-rank decomposition in language models, 2024. URL https://arxiv.org/abs/2405.06626.

Xuanfan Ni and Piji Li. A systematic evaluation of large language models for natural language generation tasks, 2024. URL https://arxiv.org/abs/2405.10251.

Ronny Paul, Himanshu Buckchash, Shantipriya Parida, and Dilip K. Prasad. Towards a more inclusive ai: Progress and perspectives in large language model training for the sámi language, 2024. URL https://arxiv.org/abs/2405.05777.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. URL https://arxiv.org/abs/2405.07101.

Mengnan Qi, Yufan Huang, Yongqiang Yao, Maoquan Wang, Bin Gu, and Neel Sundaresan. Is next token prediction sufficient for gpt? exploration on code logic comprehension, 2024. URL https://arxiv.org/abs/2404.08885.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. Large language models meet nlp: A survey, 2024a. URL https://arxiv.org/abs/2405.12819.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers, 2024b. URL https://arxiv.org/abs/2404.04925.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification in semantic space for large language models, 2024. URL `https://arxiv.org/abs/2405.13845`.

Zheyan Qu, Lu Yin, Zitong Yu, Wenbo Wang, and Xing zhang. Coursegpt-zh: an educational large language model based on knowledge distillation incorporating prompt optimization, 2024. URL `https://arxiv.org/abs/2405.04781`.

Pol G. Recasens, Yue Zhu, Chen Wang, Eun Kyung Lee, Olivier Tardieu, Alaa Youssef, Jordi Torres, and Josep Ll. Berral. Towards pareto optimal throughput in small language model serving, 2024. URL `https://arxiv.org/abs/2404.03353`.

Lorenzo Sani, Alex Iacob, Zeyu Cao, Bill Marino, Yan Gao, Tomas Paulik, Wanru Zhao, William F. Shen, Preslav Aleksandrov, Xinchi Qiu, and Nicholas D. Lane. The future of large language model pre-training is federated, 2024. URL `https://arxiv.org/abs/2405.10853`.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Rethinking llm memorization through the lens of adversarial compression, 2024. URL `https://arxiv.org/abs/2404.15146`.

Kaize Shi, Xueyao Sun, Qing Li, and Guandong Xu. Compressing long context for enhancing rag with amr-based concept distillation, 2024. URL `https://arxiv.org/abs/2405.03085`.

Noah Y. Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models, 2024. URL `https://arxiv.org/abs/2404.03189`.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, Shengwu Xiong, and Seid Muhie Yimam. Ethiollm: Multilingual large language models for ethiopian languages with task evaluation, 2024. URL `https://arxiv.org/abs/2403.13737`.

Georgy Tyukin. Enhancing inference efficiency of large language models: Investigating optimization strategies and architectural innovations, 2024. URL https://arxiv.org/abs/2404.05741.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks, 2024a. URL https://arxiv.org/abs/2404.06480.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. A user-centric benchmark for evaluating large language models, 2024b. URL https://arxiv.org/abs/2404.13940.

Bing-Yue Wu, Utsav Sharma, Sai Rahul Dhanvi Kankipati, Ajay Yadav, Bintu Kappil George, Sai Ritish Guntupalli, Austin Rovinski, and Vidya A. Chhabria. Eda corpus: A large language model dataset for enhanced interaction with openroad, 2024. URL https://arxiv.org/abs/2405.06676.

Zhenning Yang, Ryan Krawec, and Liang-Yuan Wu. Adversarial attacks and defense for conversation entailment task, 2024. URL https://arxiv.org/abs/2405.00289.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Large language models are good spontaneous multilingual learners: Is the multilingual annotated data necessary?, 2024a. URL https://arxiv.org/abs/2405.13816.

Xiang Zhang, Khatoon Khedri, and Reza Rawassizadeh. Can llms substitute sql? comparing resource utilization of querying llms versus traditional relational databases, 2024b. URL https://arxiv.org/abs/2404.08727.

Xinxin Zheng, Feihu Che, Jinyang Wu, Shuai Zhang, Shuai Nie, Kang Liu, and Jianhua Tao. Ks-llm: Knowledge selection of large language models with evidence document for question answering, 2024. URL https://arxiv.org/abs/2404.15660.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. Benchmarking large language models on cflue – a chinese financial language understanding evaluation dataset, 2024. URL https://arxiv.org/abs/2405.10542.

Longwei Zou, Qingyang Wang, Han Zhao, Jiangang Kong, Yi Yang, and Yangdong Deng. Cqil: Inference latency optimization with concurrent computation of quasi-independent layers, 2024. URL https://arxiv.org/abs/2404.06709.