Assessing LLMs for High Stakes Applications

Shannon K. Gallagher
Jasmine Ratchford
Tyler Brooks
Bryan Brown
Eric Heim
Scott McMillan
William R. Nichols
Swati Rallapalli
Carol Smith
Nathan VanHoudnos
Nick Winski
Andrew O. Mellinger
Software Engineering Institute,
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

Large Language Models (LLMs) promise strategic benefit for numerous application domains. The current state-of-the-art in LLMs, however, lacks the trust, security, and reliability which prohibits their use in high stakes applications. To address this, our work investigated the challenges of developing, deploying, and assessing LLMs within a specific high stakes application, intelligence reporting workflows. We identified the following challenges that need to be addressed before LLMs can be used in high stakes applications: (1) challenges with unverified data and data leakage, (2) challenges with fine tuning and inference at scale, and (3) challenges in reproducibility and assessment of LLMs. We argue that researchers should prioritize test and assessment metrics, as better metrics will lead to insight to further improve these LLMs.

KEYWORDS

Large language models, TEVV, metrics, scaling, HCI, trust

ACM Reference Format:
Shannon K. Gallagher, Jasmine Ratchford, Tyler Brooks, Bryan Brown, Eric Heim, Scott McMillan, William R. Nichols, Swati Rallapalli, Carol Smith, Nathan VanHoudnos, Nick Winski, and Andrew O. Mellinger. 2024. Assessing LLMs for High Stakes Applications. In *46th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3639477.3639720

1 INTRODUCTION

The Office of the Director of National Intelligence of the United States (ODNI) tasked the Software Engineering Institute to assess the viability of using large language models (LLMs) in high stakes applications such as intelligence reporting (e.g. [6]). Although LLM research has progressed rapidly in academia, government, and industry [11, 12, 16], several concerns remain unaddressed for LLMs that are intended for use in high stakes applications such as supply chain, finance, energy, and national security. Here we define a high stakes application as one that can affect people or systems in a significant manner. Other examples include creating code for defense production systems, and instructing driverless vehicles.

ICSE-SEIP '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0501-4/24/04.
https://doi.org/10.1145/3639477.3639720

Table 1: Challenges encountered in high stakes applications with LLMs and suggested research needed to address them.

Challenges	High Stakes	Research Needed
Unverified Data	Data poisoning, misinformation	Secure AI systems, human oversight
Data Leakage	PII Disclosure, safety	Safety, machine unlearning, responsible AI
Scaling training	Catastrophic forgetting	Optimization experiments, metrics
Scaling inference	Latency and computational costs	Algorithms, efficiency
Reproducibility	Traceability, accountability	Containers, seeds, versioning
Evaluation (TEVV)	Trust, security, and reliability	Metrics development, responsible AI

Failures of LLMs used in high stakes applications may exacerbate consequences and challenges in using them are magnified. We identified challenges in software engineering about trust, security, and reliability that include: assessing the quality of LLMs with regards to data; scale; and processes for Testing, Evaluation, Validation and Verification (TEVV) (see Table 1). We discuss these in detail in Section 2, and we finally suggest next steps in Section 3.

2 CHALLENGES

In this section we describe the main challenges, our approach to addressing them, and extrapolate to the larger industry community.

2.1 Challenges in Data

We suggest addressing the challenges in data through research in human oversight, securing AI systems, machine unlearning, explainable AI, and responsible AI.

In our work, we fine tuned a foundational model, as is common in the field. Note that foundational models do not store data in a typical text format, but rather represent the data through large internal weight matrices (approx. 4 GB per billion single precision weights). In general, foundational models may expose additional systems to risk because the data internally represented in them are largely unverified. One concern is that the data may be *poisoned*. Additionally, personally identifiable information (PII) or sensitive information may be *leaked* in the data via prompt engineering. Because of this, we must consider how to respond when privileged data becomes available from the LLM and if a LLM can "unlearn" [2] specific data.

Another data issue we experienced in our work is the risk involved with exposing LLMs to new, internal data. A benefit of LLMs is that they can orchestrate existing knowledge systems for information retrieval, question answering, and document summaries. However, the way information is synthesized and disseminated to the user can introduce risks, such as data leakage. To minimize

risks, human review is generally required before data can be aggregated and disseminated. While LLMs may make this process more efficient, critical oversight by subject matter experts will still be needed to mitigate risk.

The challenge of preventing data leakage via aggregation is relevant beyond the narrow focus of government organization sensitive and classified information. This issue is relevant for any situation where a LLM has access to stratified, sensitive information, such as in hospitals where various staff may have different levels of approved access to health records or PII. Thus, it is critical that software engineering practices are developed and measured to account for these edge cases and data security.

2.2 Challenges in Scale

We suggest addressing the challenges in scale by improving upon existing optimization methods and better quantifying their efficiency.

In addition to challenges with data, LLMs also require extensive engineering to work at scale, including optimizations in training and inference. For example, LLMs memory requirements can result in even smaller models' memory requirement far exceeding the best of breed GPUs available. Scaling technologies such as ZeRO [14] provide alternative methods for training LLMs via vertical scaling (i.e. splitting data across nodes) and horizontal scaling (i.e. splitting the model across nodes) and are demonstrated through implementations such as DeepSpeed [15]. Additionally, technologies such as LoRA [7] (with a sample implementation called PEFT [3]) can reduce trainable parameters by a factor of 10 or more resulting in even larger models being trained on the same hardware. By combining these methods, we estimate that a 7B parameter model can be fine-tuned with 10,000 documents using only 8 x 40G NVIDIA A100s in under 24 hours [4].

However, these scaling techniques come with significant levels of code complexity and risk, and the broader effects remain largely unexplored. In brief, LoRA based fine tuning adds new knowledge by training only a small percentage (0.1% - 10%) of model parameters. Empirically, LoRA-based fine tuning works well for many tasks (e.g. [5, 17]) and allows organizations to conduct experiments that would otherwise be too expensive. Yet, the side effects of LoRA based fine tuning are largely unknown, such as whether important information is being *forgotten*.

Finally, *inference latency* and *computational cost* are key limiters in applying LLMs to high stakes applications. In our work, responses often took over a minute to complete on standard GPU-enabled machines. This limits the scenarios in which LLMs can be used, may preclude LLMs on smaller, low-powered machines, and becomes a barrier for end users. To improve this, companies are developing special-purpose hardware to support efficient computation on smaller, quantized data types (as small as 2-bits) [10]. As a result, we need to research the effects of extreme quantization on existing LLMs. With further similar advances in hardware, LLMs may be able to run on smaller systems, requiring less power, while also decreasing response times.

2.3 Challenges in TEVV

We suggest addressing the challenges in TEVV with improving responsible AI development practices, AI reproducibility, and to prioritize metric development.

Although challenges in data and scaling are considerable, the most compelling challenge is the development and integration of processes to improve TEVV. Current metrics for LLMs include quantitative scores like BLEU [13] and ROUGE [9], more holistic assessments for accuracy, bias, and fairness such as HELM [1], and many leader board style comparisons from custom tests [18]. However, there are significant gaps in methods for assessing an LLM across a broader spectrum of quality attributes. There is no single comprehensive assessment that indicates if a LLM is trustworthy, secure, and reliable. Even custom assessments are situational or context dependent. For example, a LLM for intelligence reporting could only be fully implemented after rigorous testing with expert input. Thus, a portion of our work focused on TEVV via reproducibility and metric development.

The first challenge is experimental reproducibility, a task of complete environmental control. For software engineering, it involves configuration management such as complete platform standardization and control (e.g. containers), the same tools (e.g. version controls) and the same processes such as scripts. Reproducibility in LLM development also requires the removal of all randomness for data loading, transformation, model construction, and training. This becomes especially challenging in multi-threaded, multi-process and multi-node environments. Missed configuration details (e.g. worker thread ordering) can lead to unexpected variability. To help address this problem, we segment our activities and perform operations in separate verifiable stages when possible.

The second challenge is metric development to better assess LLMs, not only in data and and scale benchmarks but also those in trustworthiness, security, and reliability. We rely on both fine tuning and orchestration to better justify our model results, often requiring attribution for answers. Going forward, organizations will likely have to work with subject matter experts to design experiments that compare LLM responses to expert responses. Performance metrics for explainability and uncertainty quantification of LLMs will be important to support users in gaining calibrated trust [8] and becoming productive users of LLMs .

3 NEXT STEPS

LLMs are a transformative technology that have the potential to change how analytical research is performed. However, LLMs integrated into workflows for high stakes applications need improved software engineering processes to ensure they are trustworthy, secure, and reliable. To this end, we suggest addressing three critical challenges: data management, scaling, and TEVV. Of these three, developing new processes for TEVV will provide the most immediate impact. This is because TEVV addresses the largest open challenge in this field: determining how well suited a LLM is for a given application; how and if its performance is improving after fine-tuning; and how to best combine fine tuning with post-processing, prompt engineering, and APIs through orchestration.

4 ACKNOWLEDGMENTS

Copyright 2024 ACM. This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. DM24-0042

REFERENCES

[1] Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences* 1525, 1 (2023), 140–146. https://doi.org/10.1111/nyas.15007 arXiv:https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/nyas.15007

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 141–159. https://doi.org/10.1109/SP40001.2021.00019

[3] Hugging Face. 2023. PEFT. Available online at https://huggingface.co/docs/peft/index.

[4] Shannon K. Gallagher, Andrew O. Mellinger, Jasmine Ratchford, Nick Winski, Tyler Brooks, Eric Heim, Nathan VanHoudnos, Swati Rallapalli, William R. Nichols, Bryan Brown, Angel McDowell, and Hollen Barmer. 2023. A Retrospective in Engineering Large Language Models for National Security. Available online at https://insights.sei.cmu.edu/library/a-retrospective-in-engineering-large-language-models-for-national-security/.

[5] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv:2304.08247 [cs.CL]

[6] Richards J Heuer. 1999. *Psychology of intelligence analysis*. Center for the Study of Intelligence.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[8] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 PMID: 15151155.

[9] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[10] D. S. Modha, F. Akopyan, A. Andreopoulos, R. Appuswamy, J. V. Arthur, A. S. Cassidy, P. Datta, M. V. DeBole, S. K. Esser, C. Otero, J. Sawada, B. Taba, A. Amir, D. Bablani, P. J. Carlson, M. D. Flickner, R. Gandhasri, G. J. Garreau, M. Ito, J. L. Klamo, J. A. Kusnitz, N. J. McClatchey, J. L. McKinstry, Y. Nakamura, T. K. Nayak, W. P. Risk, K. Schleupen, B. Shaw, J. Sivagnaname, D. F. Smith, I. Terrizzano, and T. Ueda. 2023. IBM NorthPole Neural Inference Machine. In *2023 IEEE Hot Chips 35 Symposium (HCS)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–58. https://doi.org/10.1109/HCS59251.2023.10254721

[11] U.S. Department of Defense. 2023. DOD Announces Establishment of Generative AI Task Force. Available online at https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/.

[12] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[14] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. arXiv:1910.02054 [cs.LG]

[15] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3505–3506.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] Eric J. Wang. 2023. Alpaca-LoRA. Available online at https://github.com/tloen/alpaca-lora.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]

Received 6 October 2023; revised 10 January 2024