An experience report on finetuning LLMs for toxicologic pathology applications in pharmaceutical drug discovery

Arijit Patra, Peter Hall, Phil Scordis

UCB Pharma UK, 208 Bath Road, Slough SL1 3WE, United Kingdom
Arijit.Patra@ucb.com

Abstract. The process of drug development is a complex undertaking involving many years of research and validation. A critical part of the process lies in non-clinical safety evaluation where animal models are used to validate the toxicity, efficacy and safety profiles of investigational new drugs. This is a time-consuming process with the requirement of significant expertise, particularly when dealing with large volumes of information in the toxicologic pathology stages. With the emergence of large foundation models that can represent large scale information, it was deemed interesting to explore whether such models could be adapted to bespoke datasets that are contained within repositories of sensitive information around drug safety evaluation datasets. We explored the feasibility of finetuning open source LLMs using a specially curated toxicologic pathology dataset as an exemplar of approach, in a regulated environment of a pharmaceutical organisation, and studied the adoption of such domain specific language models and chatbots derived from them. Preliminary observations point to the challenges in driving adoption emanating from the issues around hallucination and insufficient context for specific domains where accuracy of information is critical. The abilities to perform knowledge discovery within large sets of digitised reports of past safety studies through a RAG pipeline, however proved to be of significant utility to the internal safety and pathology teams, leading to our understanding that such applications may be the pathway to widespread LLM adoption in the pharmaceutical industry and allied domains.

1 Introduction

In recent years, the adoption of transformer based large foundation models trained on significant volumes of multimodal datasets has gained prominence. The emergence of large language models, foundation models that are trained on large corpora of text resources across thematic areas, has spurred the development of a slew of language based applications that have become popular due to the ability to simulate near-human conversation and foster knowledge discovery applications in a range of use cases [1]. In this abstract, we seek to provide a glimpse into our experience of developing chat applications for non-clinical safety processes such as toxicologic pathology in the pharmaceutical drug development pipeline in a major biopharma organisation.

Pharmaceutical companies have vast libraries of bespoke safety data (toxicologic pathology and other data like toxicokinetic data, biodistribution, etc) that are stored in an easily accessible standardised format (PDF text files). Corporate knowledge is distributed across departments and is usually incomplete due to the size of the databases (a single study report may contain >1000 pages). The fragmented nature of corporate knowledge creates information islands and potential blind spots. In addition, staff attrition may lead to knowledge loss. LLMs offer the potential to fine tune on bespoke corporate knowledge and act as an easily searchable interface for knowledge retrieval. LLMs can ingest vast sets of data generated across multiple disciplines, project programs and datasets. LLMs may therefore recognise more nuanced patterns in data and generate deeper insights than human pathologists on a range of complex animal model evaluation scenarios.

2 Methodology

Datasets. We curated an instruction dataset of 100 items comprising of safety related queries and expert responses for a set of previously explored drug candidates that had advanced to the preclinical stages of the pipeline. The datasets are currently being expanded to include information about molecular chemistry and appropriate DMPK data where appropriate. We tokenized the dataset using Huggingface's AutoTokenizer, with appropriate left padding applied so that the output responses are generated as continuations of the input prompts without any unintended gaps. Given the specific nature of the application as toxicologic pathology is a specialist field within the drug development world, we proceed to finetune a Mistral 7B v0.1 LLM available through HuggingFace using parameter efficient finetuning (PEFT) [2], which allows us to freeze a proportion of the model parameters and enables to train a small percentage of the weights, thereby supporting low data situations to efficiently finetune the LLM on our domain dataset. The rationale for finetuning on a large LLM is the possibility of establishing larger context for the queries pertinent to toxicity studies within larger open world knowledge that are implicit in foundation models trained on web-scale data such as the Mistral family of architectures [3]. The model is trained for 1000 epochs, with checkpoints saved at every 50 steps, on 4 H100 GPUs available through an AzureML

subscription. Subsequently, a Retrieval Augmented Generation (RAG) [4] based pipeline was setup with the fine-tuned model for allowing knowledge discovery within existing internal safety study reports.

Non-technical aspects. An important aspect of building LLM applications that are geared towards industrial use cases in highly regulated environments such as ours is to secure appropriate levels of engagement with the intended end users, such as safety specialists and toxicologic pathologists. Additionally, given the constraints of security and compliance, internal audit and information risk management teams were involved in enabling the data curation phase as well as the quality assurance stages over the data curation and experimental design. Particularly, the technical aspect of finetuning the model was followed by an examination of model outputs by in-house pathology teams for quality validation of model responses to a range of toxicity-related queries.

Evaluations. The finetuned model endpoints are to be exposed through a custom user interface to a set of three pathologists (to control for inter-observer variability), to be evaluated using a questionnaire designed using the best practice guidelines from the Society of Toxicologic Pathologists [5] and related resources. The quality of responses are to be rated on a scale of 1 to 5 (1 – poor; 5- excellent) in terms of the scientific validity, substantive quality and clinical utility. This is to be done to account for the possibility of hallucinations leading to scientifically inappropriate content from creeping in as a result of the original base model being trained on very large-scale public and private datasets.

3 Observations, Challenges, future directions

Adoption of machine learning within the pharmaceutical industry is still nascent. There is enormous potential to harness the power of LLMs but so far little progress has been made. Multiple barriers to adoption exist due to technical and structural reasons. LLMs lack accuracy sometimes so human resource need to be expended to check the results which reduces their effectiveness. This might be especially important for smaller pharma companies where FTE (full time employee) and budget is more critical. Overall, the presented work on building finetuned LLMs for toxicologic pathology and non-clinical safety is a representation of potentially one of the first such efforts in the pharmaceutical industry and thus may be an opportunity to learn and develop best practices for such initiatives. Our current focus is on improving the quality of training datasets and publicly available versions of the same. Lastly, it is noted that the role of the human factors involved in the deployment and adoption stages need to be appreciated and incorporated in the conceptualisation of such platforms in commercial drug development processes.

References

1. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
2. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., ... & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3), 220-235.
3. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
5. Perry, R., Farris, G., Bienvenu, J. G., ... & Short, B. (2013). Society of Toxicologic Pathology position paper on best practices on recovery studies: the role of the anatomic pathologist. Toxicologic pathology, 41(8), 1159-1169.