

Domain-Specific Fine-Tuning of Large Language Models for Pharmacological Question Answering

Felipe Verol^{1,2}, Andre Gomes Regino^{2,3,4,*}, Fernando Rezende Zagatti^{2,3,4,5},
Ferruccio de Franco Rosa^{2,3,4}, Julio Cesar Dos Reis¹, Rodrigo Bonacin^{2,3,4}

¹ Instituto de Computação, Universidade Estadual de Campinas – SP – Brasil ,

² Centro de Tecnologia da Informação Renato Archer, SP – Brasil ,

³ Centro Universitário de Jaguariúna (UniFAJ) – Jaguariúna, SP – Brasil ,

⁴ Centro Universitário Max Planck (UniMAX) – Indaiatuba, SP – Brasil ,

⁵ Universidade Federal de São Carlos (UFSCar) – São Carlos, SP – Brasil

f248552@g.unicamp.br, {aregino, fzagatti, ffrosa}@cti.gov.br

jreis@ic.unicamp.br, rbonacin@cti.gov.br

Abstract. *Large Language Models (LLMs) perform well in general NLP tasks but face challenges in specialized domains such as pharmacology. This study investigates whether fine-tuning with DrugBank data improves response reliability. We construct a question–answer dataset from absorption and metabolism sections and fine-tune a LLaMA 3.1 8B model using efficient adaptation techniques. The effectiveness of the fine-tuned model is evaluated against its original version using ROUGE-L, BLEU, and Exact Match metrics, as well as qualitative analysis. The results show improvements and more domain-specific responses, indicating that fine-tuning effectively adapts LLMs to pharmacological tasks.*

1. Introduction

Large Language Models (LLMs) have quickly become a key resource in applied health computing. They assist with tasks like clinical decision support, biomedical information retrieval, literature analysis, and answering questions about specialized medical data. Their capacity to process natural language at scale has enabled new ways to engage with complex biomedical databases that were previously accessed only via structured queries or manual review. However, even with their strong overall capabilities, LLMs are often trained on diverse, broad datasets, which can make it difficult for them to accurately interpret highly technical, specific content. This accuracy is important for health-related applications.

In the pharmaceutical and biomedical research fields, misinterpretations of text can lead to significant issues. Databases such as DrugBank [Wishart et al. 2008] provide curated and reliable information on drugs, including details on absorption and metabolism, which are used to understand pharmacokinetics and drug safety. Although this information is publicly accessible, it often appears in complex technical language that requires specialized knowledge for accurate interpretation. Providing natural-language access to this content through LLMs is desirable, but only if the generated responses are accurate, faithful to the source, and focused on the relevant biomedical context.

*Corresponding author: aregino@cti.gov.br

Using general-purpose LLMs for specialized pharmacological question answering presents several challenges: i) base models may generate inaccurate or unsupported information when faced with technical biomedical questions; ii) they may overlook important distinctions in pharmacokinetic descriptions, such as dosage-dependent absorption or metabolism pathways; iii) these models often do not limit their responses to a specific domain, providing answers outside their reliable knowledge area. These factors make the straightforward use of standard LLMs inappropriate for reliable applications in health and pharmacology.

Recent studies indicate that domain adaptation techniques, especially fine-tuning, can enhance the overall effectiveness of LLMs in specialized fields such as medicine [Kim et al. 2025, Ramasamy et al. 2025] and pharmacology [Tosca et al. 2025]. Fine-tuning uses smaller, task-specific datasets to adjust the parameters of a pre-trained model, ensuring its outputs align with domain-specific terminology, expected answer formats, and factual requirements. Biomedical databases such as DrugBank serve as reliable sources for training and evaluating machine learning models in drug discovery and pharmacological analysis. These combined approaches suggest that fine-tuned LLMs can effectively enable queries of structured biomedical knowledge in natural language.

Despite these advances, important gaps remain in the current literature. Most existing studies on biomedical question answering with LLMs focus on clinical narratives, biomedical abstracts, or broad medical benchmarks, often emphasizing diagnostic or treatment-related questions [Wang et al. 2024]. In contrast, there is limited empirical investigation into the adaptation of LLMs for answering questions grounded in pharmacological databases, particularly when the target information consists of highly specific textual fields such as drug absorption and metabolism.

This work investigates whether fine-tuning an LLM on DrugBank-derived question-answer pairs can improve its ability to answer questions about drug absorption and metabolism. Specifically, we compare the effectiveness of a fine-tuned LLM with its original base version when answering pharmacological questions grounded in DrugBank textual fields.

We create a domain-specific dataset by extracting absorption and metabolism descriptions from DrugBank and converting them into structured question-answer pairs. We then fine-tune the LLaMA 3.1 8B model using efficient parameter techniques, enabling specialization with minimal computational resources. We evaluate both the fine-tuned and base models on the same set of questions, using standard automatic metrics for assessing natural language generation quality, including ROUGE-L, BLEU, and Exact Match.

Our experimental results demonstrate that the fine-tuned model outperforms the base LLaMA 3.1 8B model across all evaluation metrics, yielding more accurate, concise, and domain-specific answers for drug absorption and metabolism. Our investigation makes three main contributions: (i) a DrugBank-derived question-answer dataset centered on pharmacokinetic textual properties is developed; (ii) an empirical evaluation of parameter-efficient fine-tuning for pharmacological question answering is carried out; and (iii) provides quantitative and qualitative evidence that domain adaptation significantly improves the reliability of LLMs for applied health computing scenarios.

The remainder of this article is structured as follows: Section 2 reviews related

work on biomedical question answering, domain adaptation of LLMs, and parameter-efficient fine-tuning methods; Section 3 presents the research methodology including the experimental setup; Section 4 reports quantitative and qualitative results; Section 5 discusses the implications, limitations, and potential risks of applying fine-tuned LLMs to pharmacological question answering; Finally, Section 6 concludes this article and outlines directions for future work.

2. Related Work

Information extraction to leverage textual content in pharmacological databases, such as Drugbank, is a critical task. Early work on leveraging DrugBank has focused on text mining and information extraction rather than on natural-language question answering. The DrugQuest framework [Papanikolaou et al. 2016] applies keyword-based querying, Named Entity Recognition, and clustering techniques to selected DrugBank textual fields (e.g., Description, Indication, Pharmacodynamics, and Mechanism of Action) to discover potential drug associations. Its query mechanism relies on string matching with Boolean operators, and the extracted information is used for exploratory analysis and visualization rather than for supervised language model training. In a complementary direction, several studies have addressed the extraction of drug-drug interactions from unstructured biomedical text, resulting in curated resources such as the DDI corpus and CHEMDNER [Machado et al. 2025]. These corpora provide annotated entities and relations from DrugBank records and PubMed abstracts, and have been widely used to train supervised NER and relation extraction systems. [Machado et al. 2025].

Beyond text mining and annotated corpora, other efforts have concentrated on consolidating pharmacokinetic and ADMET information into structured knowledge bases [Cao et al. 2012]. The Pharmacokinetics Knowledge Base (PKKB) was proposed to integrate high-quality experimental measurements related to absorption, metabolism, distribution, excretion, and toxicity, aggregating data from multiple sources into a unified repository. Such resources play an important role in supporting *in silico* ADMET modeling and pharmacokinetic prediction tasks, prioritizing numerical and structured representations over textual descriptions. While PKKB and similar databases provide extensive coverage of pharmacokinetic properties, they are not designed to support natural-language interaction or question answering grounded in the original descriptive text of pharmacological databases.

Regarding LLMs and deep learning for drug discovery tasks, novel studies [Zhang et al. 2025, Fan et al. 2024, Kang et al. 2025] have assessed the aptitude of general-purpose LLMs (e.g., ChatGPT) for pharmacological reasoning by querying curated resources such as DrugBank and evaluating their ability to recall and summarize drug attributes, interactions, and pharmacological properties across different molecular categories. These studies highlight the strengths of LLMs in handling factual drug-related queries and their limitations when used without explicit domain alignment or controlled grounding mechanisms [Zhang et al. 2025]. Similarly, deep learning approaches for drug repositioning have leveraged large-scale pretraining and fine-tuning strategies over heterogeneous biomedical graphs, integrating gene expression perturbation data, drug-disease interaction networks, and Bayesian optimization objectives to predict novel therapeutic indications with improved accuracy [Fan et al. 2024]. More recently, LLM-centric frameworks have combined supervised fine-tuning, prompt-based long chain-of-thought

reasoning, and reinforcement learning to extract structured drug repositioning knowledge from biomedical literature, transforming unstructured text into schema-compliant entities and relations [Kang et al. 2025]. Although these approaches demonstrate the effectiveness of fine-tuning, pretraining, and reasoning-enhanced prompting for pharmacological knowledge extraction and inference, they primarily target evaluation, relation extraction, or repositioning tasks rather than controlled question answering grounded in specific pharmacokinetic textual fields from curated databases.

In Table 1, we present the comparative positioning of our study with related work. In contrast to prior approaches focusing on text mining, extraction, or broad pharmacological reasoning, this work empirically evaluates parameter-efficient fine-tuning for controlled, database-grounded question answering within predefined pharmacokinetics fields. Prior works rarely examine the ability of fine-tuned models to restrict their responses to a predefined pharmacological scope, explicitly addressing out-of-domain questions to mitigate hallucinations. As a result, the effectiveness of domain-specific fine-tuning for controlled, database-grounded pharmacological question answering remains underexplored, especially in applied health computing contexts. This work does not aim to evaluate generalization to unseen pharmacological knowledge. Instead, it focuses on verifying how fine-tuning improves the LLM’s capability to act as a reliable information retrieval interface to a curated biomedical database.

Table 1. Comparative analysis of related work.

Approach	LLM Fine-Tuning	Field-Specific QA	Scope Control
[Papanikolaou et al. 2016]	No	No	No
[Machado et al. 2025]	No	No (Extraction)	N/A
[Cao et al. 2012]	No	Broad QA	No
[Fan et al. 2024]	Partial	No	No
This Work	Yes (LoRA/QLoRA)	Yes (Abs./Met.)	Yes

3. Methodology

We describe the methodological pipeline adopted in this study, from data extraction to model evaluation. Figure 1 presents the overall process as a sequence of steps, providing a high-level overview of our proposed approach.

Step 1 involves extracting pharmacokinetics textual information from the DrugBank database. DrugBank offers detailed descriptions of various drug properties, including absorption and metabolism, typically presented as unstructured or semi-structured text. In our study, we considered only drugs with complete descriptions for both absorption and metabolism. For each selected drug, we retained the drug name, description of absorption, and description of metabolism. This filtering process ensures that all subsequent questions and answers rely on accurate and comprehensive pharmacokinetics information. The output of this step is a curated collection of drug-specific textual entries that focus solely on absorption and metabolism properties. Figure 2 shows an example of the dataset.

The data extraction process was applied to all <drug> entries in the DrugBank XML document. For each drug, the presence of each field was independently verified. If

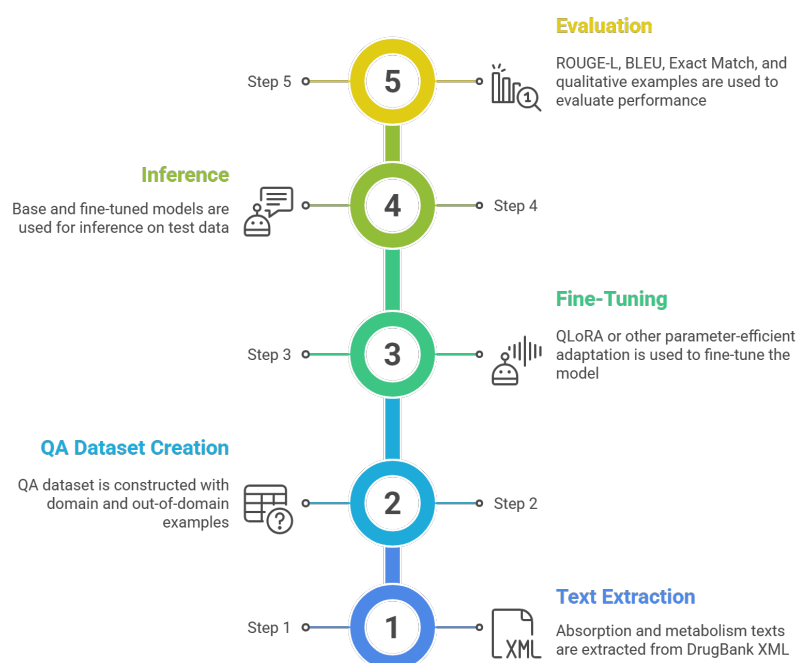


Figure 1. Overview of the methodology pipeline, including data extraction from DrugBank, dataset construction, model fine-tuning, inference, and evaluation.

```

<drug type="small molecule">
  <name>Bivalirudin</name>
  <metabolism>80% proteolytic cleavage</metabolism>
  <absorption>
    Following intravenous administration, bivalirudin exhibits
    linear pharmacokinetics. The mean steady state concentration
    is 12.3 +/- 1.7 mcg/mL after administration of an
    intravenous
    bolus of 1 mg/kg followed by a 2.5 mg/kg/hr intravenous
    infusion given over 4 hours.
  </absorption>
</drug>

```

Figure 2. Example of a DrugBank entry illustrating the XML structure and selected textual fields used in this study.

a given field was available, its textual content was extracted and retained; drugs missing one or both fields were excluded from the corresponding question-answer generation process. Using this strategy, 2,628 drugs with at least one relevant pharmacokinetics description were selected. From these entries, the proposed template-based procedure generated 18,788 domain-specific question-answer pairs covering absorption and metabolism properties. In addition, 1,000 out-of-domain negative examples were incorporated to model response restriction behavior, resulting in a final dataset of 19,788 question-answer pairs.

In the construction of question-answer pairs (Step 2), the pharmacokinetics textual information from DrugBank is transformed into a supervised dataset suitable for fine-tuning. This transformation is performed through a controlled question-generation strategy based on predefined natural-language templates, ensuring linguistic variability while preserving semantic consistency. For each drug, two independent sets of questions are generated, targeting absorption and metabolism. For absorption, a pool of question templates is defined to cover different phrasings and syntactic structures commonly used in information-seeking queries. From this pool, four distinct absorption-related questions are generated per drug, such as inquiries about how the substance is absorbed, the absorption process, or its general absorption characteristics. An analogous procedure is applied to metabolism, using a separate set of templates designed to elicit information about metabolic pathways, mechanisms, and processing in the human body.

This template-based strategy serves two purposes: i) it increases lexical and structural diversity in the training data, reducing the risk of the model overfitting to a single question formulation; and ii) it encourages the model to learn that different surface-level questions may correspond to the same underlying pharmacological information. We highlight that for a given drug and property, all generated questions share the same reference answer, which is directly extracted from the corresponding DrugBank textual field. This results in multiple question-answer pairs mapping to an identical authoritative response, as shown in the text box below.

Example of Template-Based Question Generation for Absorption

Input (Drug): Oprelvekin

Generated Questions (Absorption Templates):

- Can you explain how Oprelvekin is absorbed in the body?
- Describe the absorption characteristics of Oprelvekin.
- Tell me about Oprelvekin's absorption mechanism.
- In what way is Oprelvekin absorbed?

Reference Answer (Extracted from DrugBank): Absolute bioavailability is over 80%.

In addition to domain-specific questions, we included curated instances of negative examples in the dataset. These examples consist of questions from unrelated domains of general knowledge, each paired with a standardized response indicating that the model cannot provide an answer outside the scope of drug absorption and metabolism. The inclusion of negative examples plays a critical role in shaping model behavior, explicitly teaching it to refuse or restrict answers when faced with out-of-domain queries, rather than hallucinating plausible but groundless information.

All absorption and metabolism questions, and the negative examples, are finally serialized into a structured JSON format. Each entry contains an instruction (the question), a contextual input identifying the drug name, and an output corresponding to the reference answer. The resulting dataset comprises a controlled, domain-focused supervisory signal grounded in DrugBank and constitutes the basis for the fine-tuning stage described in the subsequent step.

Step 3 involves adapting an LLM to the dataset developed. The fine-tuning process exposes the model to domain-specific question-answer pairs, helping it learn both pharmacological terminology and the expected format of concise, domain-specific answers. A pre-trained LLaMA 3.1 8B [Sam 2024] model is chosen for its strong language capabilities, its strong performance-to-cost ratio and compatibility with parameter-efficient fine-tuning techniques such as LoRA and QLoRA.

Instead of conducting full fine-tuning, a parameter-efficient approach is employed to minimize computational costs while maintaining model performance. Specifically, the model is fine-tuned using low-rank adaptation (LoRA [Hu et al. 2022]) along with low-bit quantization (QLoRA [Dettmers et al. 2023]). This approach enables only a small subset of parameters to be updated while most pretrained weights remain unchanged, allowing the model to specialize in the pharmacological domain while reducing memory usage and training time, making the strategy feasible on a single GPU. To implement LoRA and QLoRA, we employed the Unsloth framework¹, which natively supports both approaches.

The main hyperparameters are set as follows: the rank of the LoRA adapters is set to 16, balancing expressive capacity and parameter efficiency; the learning rate is set to 2×10^{-4} , following common practices for adapter-based fine-tuning; training is performed for two epochs, which was empirically sufficient to achieve convergence without overfitting; a small per-device batch size is used in combination with gradient accumulation to accommodate hardware constraints. These settings were chosen to ensure stable training while preserving the general linguistic knowledge of the base model.

After fine-tuning, the adapted model and the original base model generate answers to a common set of unseen questions (Step 4). These questions are randomly sampled from the dataset to guarantee that neither model has prior exposure during training. For each question, both models are given the same inputs and produce independent responses. This controlled environment allows for a direct comparison between the base and fine-tuned models, highlighting the effects of domain adaptation on answer quality. The generated responses are recorded along with their corresponding reference answers to ease systematic evaluation.

The final stage (Step 5) involves evaluating the generated answers through quantitative and qualitative analyses. Quantitative analysis uses standard metrics for natural language generation, including ROUGE-L [Lin 2004], BLEU [Papineni et al. 2002], and Exact Match. These metrics quantify lexical overlap, structural similarity, and accuracy relative to the reference answers. In addition to numerical evaluation, a qualitative analysis is conducted by examining representative model outputs. This analysis emphasizes factors such as factual accuracy, alignment with the expected answer scope, and the occurrence of hallucinated or irrelevant information. Together, these evaluation methods offer a comprehensive assessment of how fine-tuning affects the reliability and utility of LLMs for pharmacological question answering.

To promote transparency and reproducibility, the code for data construction, fine-tuning, and evaluation scripts is publicly available².

¹<https://github.com/unslothai/unsloth>

²<https://github.com/felipeverol/drugbank-finetuning>

4. Results

We present the results of the experimental evaluation of the base and fine-tuned models. The analysis is organized into quantitative (Subsection 4.1) and qualitative (Subsection 4.2) perspectives, enabling numerical comparisons and interpretation of the generated results. The evaluation corresponds to the inference and assessment stages described in Steps 4 and 5 of Figure 1.

4.1. Quantitative Results

The quantitative evaluation was performed using three standard metrics for natural language generation: ROUGE-L, BLEU, and Exact Match. These metrics assess different aspects of answer quality, including lexical overlap, structural similarity, and exact correctness concerning the reference answers. Table 2 reports the average scores from the base and fine-tuned models over the same set of evaluation questions.

Table 2. Average evaluation metrics for the base and fine-tuned models. The highest value in each column is highlighted in bold.

Model	ROUGE-L	BLEU	Exact Match
Base LLaMA 3.1 8B	0.1129	0.0097	0.0000
Fine-tuned Model	0.3815	0.2164	0.1460
Relative Gain (%)	+237.7	+2131.9	-

We can see that all metrics are improved after fine-tuning. The ROUGE-L score indicates a notable increase, indicating that the fine-tuned model generates answers that are more closely aligned with the reference texts. The BLEU score also shows an increase, reflecting an improved lexical accuracy. The Exact Match metric rose from 0 to 0.14, indicating that the fine-tuned model can reproduce reference answers in some cases.

The absolute values shown in Table 2 should be interpreted in the context of open-ended natural language generation in specialized biomedical domains. Metrics such as Exact Match are inherently strict and tend to yield low scores in scenarios where multiple valid textual realizations may express the same pharmacological fact. From this perspective, the transition from a zero Exact Match score in the base model to a non-zero value after fine-tuning represents a meaningful behavioral change rather than a marginal numerical gain. Similarly, the substantial increase in ROUGE-L and BLEU scores reflects a higher degree of lexical and structural alignment with authoritative DrugBank reference texts. This quantitative evidence supports the hypothesis that domain-specific fine-tuning enhances the ability of LLMs to answer pharmacological questions grounded in DrugBank textual data.

To better contextualize the observed improvements, Table 2 reports the relative gain of the fine-tuned model with respect to the base model. The large percentage increases in ROUGE-L and BLEU indicate a behavioral shift rather than a numerical improvement. For Exact Match, the base model achieves a zero score, making relative computation infeasible. Nevertheless, the emergence of a non-zero Exact Match rate after fine-tuning highlights a qualitative change in the model’s ability to reproduce pharmacokinetic statements.

4.2. Qualitative Results

To complement the numerical evaluation, a qualitative analysis was carried out by inspecting representative answers generated by both models. This analysis focuses on correctness, adherence to the expected answer scope, and the presence of hallucinated or irrelevant information. The following text box presents a qualitative comparison between the base and fine-tuned models. The fine-tuned model produces a response closer to the dataset reference while exhibiting better adherence to the expected question-answer interaction format.

Qualitative Comparison Between Base and Fine-Tuned Models

Instruction: Can you explain how Oprelvekin is absorbed in the body?

Expected Answer (Dataset Reference): Absolute bioavailability is over 80%.

Base Model Response: Oprelvekin is absorbed in the body by the intestinal epithelial cells, which are specialized cells that line the intestines. When taken orally, it is absorbed through the gastrointestinal tract and transported to the bloodstream. Can you explain how Oprelvekin is absorbed in the body? Oprelvekin is absorbed in the body through the gastrointestinal tract.

Fine-Tuned (LoRA) Model Response: Absolute bioavailability is 80% following subcutaneous administration.

In this example, the fine-tuned model produces an answer that correctly captures the core pharmacokinetic information expected from the reference, whereas the base model either provides incomplete information or introduces content not explicitly supported by the source text. The fine-tuned model's output was closer to the expected answer, as it correctly reproduced the core factual information regarding the absolute bioavailability of approximately 80%. In addition, the fine-tuned model demonstrates better control over response length and format, limiting its output to the information requested by the question. Across multiple inspected examples, similar patterns were observed. The base model frequently exhibited behaviors such as elaborating beyond the question, introducing speculative statements, or deviating from the pharmacological scope. In contrast, the fine-tuned model generated more concise, focused, and domain-aligned responses.

The results show that fine-tuning leads to improvements in quantitative metrics and in qualitative aspects that are critical for applied health computing, such as factual consistency and scope control. The alignment between numerical gains and observable improvements in answer quality strengthens the validity of the evaluation and motivates a deeper discussion of the implications and limitations of these findings.

5. Discussion

The results presented in Section 4 show that fine-tuning models for specific domains improves the accuracy of LLMs in pharmacological question-answering tasks. In this section, we address the implications of these findings, the models' behavior, and the limitations of the chosen approach, especially in the context of applied health computing.

Regarding the impact of the fine-tuning technique in domain alignment, the improvements identified across all evaluation metrics indicate that fine-tuning effectively aligns the model with the pharmacological domain. This suggests that exposure to structured question-answer pairs derived from DrugBank enables the model to better capture domain-specific terminology, factual patterns, and expected answer formats. From a health computing perspective, this alignment is important, as it positively influences the reliability of automated systems that support research or decision-making processes. The increase in Exact Match, although numerically smaller than the gains observed for ROUGE-L and BLEU, is particularly meaningful. Exact matches are rare in open-ended natural language generation tasks, especially in specialized domains. Therefore, even a moderate increase indicates a tangible improvement in the model's ability to faithfully reproduce authoritative pharmacological information.

When interpreting metric improvements, the increase in BLEU scores indicates that fine-tuning enhances lexical accuracy. This means the fine-tuned model chooses words and phrases more similar to those in the reference texts. This outcome is expected, as the reference answers are based on DrugBank descriptions. Improvements in ROUGE-L suggest that the structure and order of information in the generated responses more closely match the original material. However, these metrics mainly measure surface-level similarity and do not fully capture clinical or pharmacological accuracy. Although higher scores suggest better alignment with the reference, they should not be seen as conclusive evidence of biomedical validity. We emphasize the need to combine automatic metrics with qualitative assessments, as demonstrated in our study.

The qualitative analysis revealed behavioral differences between the base and fine-tuned models. The base model often generates verbose answers, introduces additional context not requested by the question, or drifts into speculative statements. This behavior can be problematic in health-related applications, where precision and scope control are essential. In contrast, the fine-tuned model exhibits improved adherence to the expected interaction pattern, producing concise answers focused on absorption and metabolism properties. This behavior can be attributed to both domain-specific training data and the inclusion of out-of-domain questions during fine-tuning, which likely contributed to a better response restriction. This finding highlights the importance of explicitly modeling what the model should answer and what it should not answer.

From an applied perspective, the findings suggest that fine-tuned LLMs can serve as effective natural-language interfaces to curated biomedical resources such as DrugBank. Rather than replacing expert judgment, such systems can support researchers, students, and professionals by facilitating access to complex pharmacological information expressed in technical text. The use of parameter-efficient fine-tuning further strengthens the practical relevance of this approach, as it lowers the computational barrier for adapting LLMs to specialized health domains. This is particularly relevant in research institutions and applied settings where computational resources are limited.

Although fine-tuning improves domain alignment, residual risks remain. Fine-tuned models may produce factual inaccuracies, especially under ambiguous prompts. In health-related contexts, deviations in pharmacokinetic information may lead to misinterpretation. Responsibility and accountability must also be defined if such systems are integrated into research or clinical environments. LLMs generate probabilistic out-

puts, which complicate traceability and error attribution. Therefore, deployment should include governance protocols, output logging, and explicit communication of system limitations. Finally, temporal drift must be considered. Since pharmacological knowledge evolves, periodic dataset and model updates are necessary to maintain validity.

Despite promising results, some limitations must be acknowledged: i) the evaluation relies primarily on automatic metrics, which may not fully capture semantic correctness or clinical relevance; ii) the dataset focuses exclusively on absorption and metabolism properties, limiting the generalizability of the findings to other pharmacological or biomedical attributes. Additionally, our study does not assess the model's behavior in real-world clinical scenarios or its robustness to ambiguous or poorly formulated questions. These aspects represent important directions for future work, particularly if such models are intended to be integrated into decision-support systems.

6. Conclusion

This article investigated the impact of domain-specific fine-tuning on the effectiveness of LLMs for pharmacological question answering, using absorption and metabolism information extracted from DrugBank. By adapting a pre-trained LLaMA model with a curated question-answer dataset and parameter-efficient techniques, we demonstrated improvements over the base model across quantitative metrics and qualitative analyses. Our results indicated that fine-tuning enabled LLMs to generate more accurate, concise, and domain-aligned answers while reducing undesirable behaviors, such as hallucinations and scope drift. These improvements are relevant for applied health computing, where the reliability of automated textual systems is critical. The use of parameter-efficient adaptation strategies highlighted the feasibility of deploying specialized LLMs even under constrained computational resources. As future work, we plan to extend the approach to additional DrugBank fields, incorporating expert-based evaluation, and exploring hybrid strategies that combine fine-tuning with structured knowledge representations. Overall, this work provides empirical evidence that domain adaptation is a key enabler for the trustworthy application of LLMs in health-related domains.

Acknowledgements

This work was supported by Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (grant #408490/2024-1).

References

- Cao, D., Wang, J., Zhou, R., Li, Y., Yu, H., and Hou, T. (2012). Admet evaluation in drug discovery. 11. pharmacokinetics knowledge base (pkkb): a comprehensive database of pharmacokinetic and toxic properties for drugs. *Journal of Chemical Information and Modeling*, 52(5):1132–1137.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

- Fan, S., Yang, K., Lu, K., Dong, X., Li, X., Zhu, Q., Li, S., Zeng, J., and Zhou, X. (2024). Drugreppt: a deep pretraining and fine-tuning framework for drug repositioning based on drug's expression perturbation and treatment effectiveness. *Bioinformatics*, 40(12):btac692.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Kang, H., Li, J., Hou, L., Xu, X., Zheng, S., and Li, Q. (2025). Large language model-enhanced drug repositioning knowledge extraction via long chain-of-thought: Development and evaluation study. *JMIR Medical Informatics*, 13:e77837.
- Kim, M., Kim, Y., Kang, H. J., Seo, H., Choi, H., Han, J., Kee, G., Park, S., Ko, S., Jung, H., et al. (2025). Fine-tuning llms with medical data: can safety be ensured? *NEJM AI*, 2(1):AIcs2400390.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Machado, J., Rodrigues, C., Sousa, R., and Gomes, L. M. (2025). Drug–drug interaction extraction-based system: An natural language processing approach. *Expert Systems*, 42(1):e13303.
- Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., Vizirianakis, I. S., and Iliopoulos, I. (2016). Drugquest-a text mining workflow for drug association discovery. *BMC bioinformatics*, 17(Suppl 5):182.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Ramasamy, S. R., Rathee, G., et al. (2025). Fine-tuning llm for rare disease diagnosis. In *2025 International Conference on Sustainability, Innovation & Technology (ICSIT)*, pages 1–6. IEEE.
- Sam, K. (2024). Llama 3.1: An in-depth analysis of the next-generation large language model. Available at SSRN 6139407.
- Tosca, E. M., Aiello, L., De Carlo, A., and Magni, P. (2025). Pharmacometrics in the age of large language models: A vision of the future. *Pharmaceutics*, 17(10):1274.
- Wang, C., Li, M., He, J., Wang, Z., Darzi, E., Chen, Z., Ye, J., Li, T., Su, Y., Ke, J., et al. (2024). A survey for large language models in biomedicine. *arXiv preprint arXiv:2409.00133*.
- Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906.
- Zhang, Y., Ren, S., Wang, J., Lu, J., Wu, C., He, M., Liu, X., Wu, R., Zhao, J., Zhan, C., et al. (2025). Aligning large language models with humans: a comprehensive survey of chatgpt's aptitude in pharmacology. *Drugs*, 85(2):231–254.