ISSN 2007-9737

Comparing Domain-Adapted Small LLMs and Large Zero-Shot Models for Efficient Hypothesis Graph Construction in Neurodegenerative Disease Research

Cesar Torres, Claudia I. Gonzalez*, Mario Garcia

Tijuana Institute of Technology / TECNM,
Mexico

{cesartorres, cgonzalez}@tectijuana.mx, mario@tectijuana.edu.mx

Abstract. The rapid expansion of biomedical literature has made it increasingly difficult to intersect findings and uncover novel hypotheses, particularly within neurodegenerative disease research. This work presents a domain-adapted approach for hypothesis extraction and synthesis using compact large language models (Qwen3 0.6B–8B parameters). Through parameter-efficient fine-tuning on a curated corpus of open-access biomedical papers, these models work over identifying, relate, and visualize research hypotheses as interconnected knowledge graphs. Their performance is evaluated against a larger, general-purpose models in zero-shot conditions, demonstrating that smaller, specialized models can achieve comparable or superior interpretability and relevance. The study highlights the potential of lightweight, domain-focused LLMs as practical tools for accelerating discovery and improving transparency in biomedical research.

Keywords. Large language models (LLMs), biomedical research, hypothesis extraction, parameter-efficient fine-tuning, LoRA, small language models, neurodegenerative diseases.

1 Introduction

The field of natural language processing has been transformed by the rapid adoption of large language models (LLMs). Built upon the Transformer architecture [24], these models are capable of generating coherent text, performing operations assimilated to human reasoning with complex information across multiple domains, and achieving state-of-the-art results in tasks such as question answering, research, and summarization [18, 23]. Their success has also accelerated the adoption in domains that require advanced information extraction methodologies, including the biomedical sciences.

The widespread adoption of LLMs has been followed by exponential computational demands. State-of-the-art models commonly reach hundreds of billions grasping onto trillion of active parameters [4, 22, 27], demanding substantial computational resources for both training and inference.

Prior analyses estimate that training a single large model can consume gigawatt-hours of electricity [20] and rely on cloud-scale infrastructure with considerable carbon footprints [19]. These requirements limit accessibility for smaller laboratories and create challenges for reproducibility, long-term sustainability, and on-premise deployment [26].

These computational and infrastructural burdens have broad implications for scientific fields that increasingly rely on knowledge extraction. As research output grows and the demand for scalable analysis intensifies, the gap between what state-of-the-art LLMs require and what most applications can realistically support continues to extend. This disconnect is especially problematic in domains where timely synthesis of evidence directly influences scientific progress.

Beyond the computational challenges, the biomedical domain brings its own set of difficulties. Evidence synthesis is often messy:

studies report results using different formats, key measurements/reports are sometimes missing, and outcome definitions vary widely across diferent researchers. Earlier machine learning work showed that even straightforward unsupervised methods can help fill in missing information and standardize effect estimates across clinical studies, improving the consistency of systematic reviews and meta-analyses [16]. More recent efforts using publicly available LLMs for diagnostic support have demonstrated that smaller models can reach competitive performance in real clinical scenarios, although their outputs tend to vary without careful domain adaptation [21]. Together, these observations reinforce the need for structured, task-specific pipelines, especially in areas like neurodegeneration, where the literature is extensive, fragmented, and highly heterogeneous.

Biomedical research exemplifies both the opportunities and limitations of relying on such systems. The life sciences now generate more than a million publications every year [3], producing a volume of literature that exceeds the capacity of manual review. This challenge is acute in neuro degenerative disease research, where meaningful progress depends on connecting heterogeneous findings across studies focused on Alzheimer's, Parkinson's, Huntington's disease, and related disorders [13, 25]. While state-of-the-art LLMs such as GPT-4 achieve strong performance on medical benchmarks [17], their reliance on closed-source infrastructure and high inference cost poses significant barriers for institutions needing transparent, reproducible, and domain-tailored models.

These constraints have intensified interest in approaches that retain strong performance while reducing computational demands. Parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) [10] and advanced post-training quantization techniques [7, 28] allow small and medium-sized models to be adapted for specialized tasks at a fraction of the cost of full fine-tuning. As a result, Small Language Models (SLMs), typically ranging from hundreds of millions to single-digit billions of parameters have gained prominence as practical, energy-efficient alternatives for scientific and clinical applications [2, 8]. Recent medical and multilingual LLMs further demonstrate that compact architectures can match or exceed the performance of larger models when paired with domain-specific training data and optimization strategies [1, 12].

In parallel, the emergence of agentic AI systems has introduced new opportunities for orchestrating LLMs through collections of specialized, task-driven components. Multi-agent arrangements are able to break appart complex workflows, such as literature summarization, biomedical entity extraction, graph construction, and hypothesis generation, into modular sub tasks that can be assigned to smaller, fine-tuned models. Prior work shows that such single-task specialization can outperform larger general-purpose models, particularly under resource constraints [2, 9].

This paradigm aligns well with the demands of biomedical knowledge discovery, where transparency, reproducibility, and domain specificity are paramount.

Despite these trends, few studies have systematically examined how small, domain-adapted LLMs and agentic architectures can be combined to support neurodegenerative disease research.

To address this gap, this paper proposes an agentic AI framework that leverages small, task-specific LLMs fine-tuned with LoRA and quantization. We develop and evaluate a multi-stage pipeline for biomedical literature mining, covering text extraction, section summarization, entity tagging, knowledge graph construction, and structured hypothesis generation and assess the models ranging from 0.6B to 8B parameters on performance, efficiency, and applicability to neurodegenerative research. The contributions of this work are threefold.

First, we provide a detailed description of the agentic pipeline. Second, we conduct an expanded experimental study encompassing a new datasets, metrics such as BERTScore [30], and case studies in neurodegenerative disease. Third, we offer a deep examination of computational efficiency enabled by parameter-efficient fine-tuning and quantization. The remainder of the paper is organized as follows: Section 2 reviews background and related work; Section 3 describes the methodology; Section 4 presents the experimental

setup; Section 5 reports results and discussion; and Section 6 concludes with future directions.

2 State of the Art

In this section, we provide a literature review on the progress in the use of small and compact language models (SLMs) within agentic AI for biomedical applications. We also discuss parameter-efficient fine-tuning methodologies and graph-based approaches for knowledge discovery in scientific and biomedical literature.

2.1 Small Language Models and Agentic AI

Recent work has increasing recognized the capabilities of SMLs as a viable alternative that is more efficient to large-scale models for agentic systems [2], specially in domains like biomedical sciences where computational cost, privacy and deployment constraints are an important fact into consideration for deployments.

Recent systematics reviews of SLMs published between 2023-2025 demonstrates how models around 7 billion parameters are now state-of-the-art for downstream task, thanks to advances in training techniques and architecture optimizations [5, 8]. The evolution of this technologies suggest that SLMs are more than economical compromises, but legitimate, performant building blocks for AI systems.

In the biomedical domain, the newly introduced Meerkat illustrates the potential of medical-SLMs [11]. The model is designed to be lightweight and sport reasoning capabilities relevant to clinical tasks, a promising step towards real applications under real-world hardware following privacy constraints.

In addition, compact models show their value over classificcaiton task in constrained resources. Recent work Lightweight Baselines for Medical Abstract Classification: DistilBERT with Cross-Entropy as a Strong Default demosntrates a distilled encoder finetuned with Cross-Entropy loss can provide a solid performance over medical-abstract classifcation while requiring a lower parameter count than the base models [14].

The advances suggest an emerging paradigm, where the multiple use cases in the biomedical domain are not defaulting over large models for biomedical or agentic practices. Particularity when tasks are domain-specific, repetitive and resource constrained. Such approaches enable lower latency, more private and cost effective AI systems, which are attractive to healthcare and clinical settings.

2.2 LoRA and Related Methods for Parameter-Efficient Fine-Tuning

An essential enabler for applying small language models (SLMs) in specialized domains is the continued advancement of parameter-efficient fine-tuning (PEFT) methods. These approaches allow pre-trained models to be adapted to new tasks while updating only a small subset of parameters, significantly reducing computational cost and memory requirements. Low-Rank Adaptation (LoRA) [10] remains one of the most widely adopted techniques, and recent extensions such as QLoRA [6], DoRA [15], and AdaLoRA [29] further improve efficiency, stability, and weight allocation. These developments have made it feasible to fine-tune compact models for domain-specific biomedical tasks even on limited hardware, enabling practical deployment in real-world clinical and scientific settings.

In standard fine-tuning, a pre-trained weight matrix W_0 in a transformer layer is fully updated with an additional parameter matrix ΔW:

$$W = W_0 + \Delta W. \tag{1}$$

This approach is costly since ΔW has the same dimensionality as W_0. LoRA addresses this by constraining ΔW to be low-rank:

$$\Delta W = W_B W_A, \tag{2}$$

where $W_A \in \mathbb{R}^{r \times d_{\text{in}}}$, $W_B \in \mathbb{R}^{d_{\text{out}} \times r}$, and $r \ll \min(d_{\text{in}}, d_{\text{out}})$. In this formulation, only the small matrices W_A and W_B are trained, while the original weights W_0 remain frozen. This greatly reduces the number of trainable parameters, accelerates training, and lowers GPU

memory usage, making LoRA particularly suitable for resource-constrained environments.

Beyond LoRA, additional PEFT techniques have been proposed, including quantization [7, 28], adapter modules, and prefix-tuning. These methods share the same goal: reducing computational cost while maintaining competitive performance.

In biomedical domains, such approaches are increasingly applied to clinical NLP, medical question answering, and multimodal applications, where energy-efficient adaptation is critical for on-device deployments.

2.3 Graph-based Literature Mining and Similarity Metrics

This work relies on constructing hypothesis graphs from scientific literature. Entities are represented as nodes and their relationships as weighted edges, which makes it possible to capture the latent structure of biomedical knowledge and uncover new connections. The following similarity and clustering techniques are commonly applied:

Cosine similarity measures the semantic similarity between embedding vectors of entities or text fragments:

$$\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}. \tag{3}$$

Ontology-based similarity, such as the Information Content (IC) approach, provides a knowledge-driven metric. The Lin similarity is defined as:

$$\mathrm{sim}_{\mathrm{Lim}}(c_1, c_2) = \frac{2\, IC(\mathrm{LCS}(c_1, c_2))}{IC(c_1) + IC(c_2)}, \tag{4}$$

where c_1, c_2 are ontology concepts, $\mathrm{LCS}(c_1, c_2)$ is their lowest common subsumer, and $IC(x)$ is the information content of concept x.

Graph modularity, frequently optimized using the Louvain method, enables the identification of communities or clusters of entities:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) 1[g_i = g_j], \tag{5}$$

where A_{ij} is an entry in the adjacency matrix, k_i is the degree of node i, m is the total number of edges, and $1[g_i = g_j]$ is an indicator function that equals 1 when nodes i and j belong to the same community.

3 Methodology

Here we will describe the proposed pipeline to perform the training of the models, implementing the hypothesis graph generation extracted from the biomedical literature.

The workflow contains the generation of the dataset, extraction, summarization, entity generation, for the multiple sections in the paper, as well as the hypothesis construction and model adaptation. The overall process is illustrated in Figure 1.

3.1 Dataset Creation

To support our experiments, we collected 800 open-access biomedical papers from Pub Med Central, filtered for relevance to neurodegenerative disease. We designed a multi-agent processing pipeline in which each stage is handled by an autonomous agent responsible for a specific transformation of the data.

Raw PMC articles were converted into JSONL format and decomposed into canonical scientific sections (*Introduction*, *Methods*, *Results*, *Discussion*, etc.).

This pipeline produced two datasets:

— **Flat corpus (pre-training)**: ∼50,000 normalized paragraphs extracted from 200 papers, used for domain-adaptive pretraining.

— **Sectioned corpus (fine-tuning)**: 600 papers processed into structured section entries enriched with summaries, entities, and hypotheses for supervised fine-tuning.

3.2 Agentic Section Processing Pipeline

Dataset creation is carried out through a sequence of specialized agents, where each agent transforms the output of the previous one:

— **Section Extraction Agent**
Consumes raw full-text articles, normalizes XML/HTML structure, spliting them into standardized scientific sections while preserving metadata such as PMCIDs and disease labels. This step produces the first structured representation of each paper.

— **Summarization Agent**
Uses a LLM (Qwen3-32B) [27] to generate concise, technically grounded summaries for every section. The prompting strategy enforces factuality by requiring the model to extract mechanistic statements, highlight major findings, and cite explicit evidence sentences.

— **Entity Tagging Agent and Graph Creation**
Processes each summary with biomedical NER models to extract genes, proteins, pathways, diseases, and other domain-relevant concepts. This attaches a structured set of biological entities to each section, then this tags get passed to an LLM to perform a knowledge graph creation with the weights of the tags.

— **Hypothesis Generation Agent**
Takes the summaries and their associated entities and produces candidate mechanistic hypotheses at the paper level. This agent uses a two-pass approach: a ranking pass to identify salient evidence or entity combinations, followed by a controlled generation pass that formulates grounded, technically accurate hypotheses.

Together, these agents form a coordinated, end-to-end pipeline (Figure 1) that converts unstructured biomedical literature into a dataset.

The modular design ensures that each stage is interpretable, verifiable, and replaceable as improved models and tools become available.

4 Experimental Setup

This section describes the practical aspects of adapting and validating models, including pretraining, fine-tuning, evaluation metrics, and hardware resources. The model training procedure is illustrated in Figure 2.

4.1 Training Procedure

Model training is carried out in two stages, reflecting the structure of the constructed datasets and the design of our agentic pipeline (Figure 2). The goal of this procedure is to (1) endow the model with broad biomedical domain knowledge, and (2) specialize it for mechanistic hypothesis generation using the structured outputs of our agents.

Stage 1: Domain-Adaptive Pre-training The first stage adapts the base LLM to scientific language and biomedical discourse. We perform continued pre-training using the flat corpus, consisting of approximately 50,000 normalized paragraphs derived from complete, unprocessed PMC articles. Because these paragraphs preserve natural scientific flow, linguistic variability, and contextual richness, this step equips the model with improved terminology grounding, paragraph-level coherence, and familiarity with biomedical writing conventions. No instruction tuning or supervision is used at this stage; the objective is purely next-token prediction over the domain corpus.

Stage 2: Supervised Fine-Tuning for Hypothesis Generation In the second stage, the model is trained on the structured dataset produced by the proposed agent pipeline. Each training instance consists of a section-level summary, its associated biomedical entity graph, the local evidence extracted by the agents, and the final hypothesis generated at the paper level. This sectioned corpus enables the model to learn explicit mappings between evidence, biological concepts, and mechanistic reasoning patterns.

To align the model with the structure of the agent-generated dataset, we adopt a dual-supervision strategy. The training objective

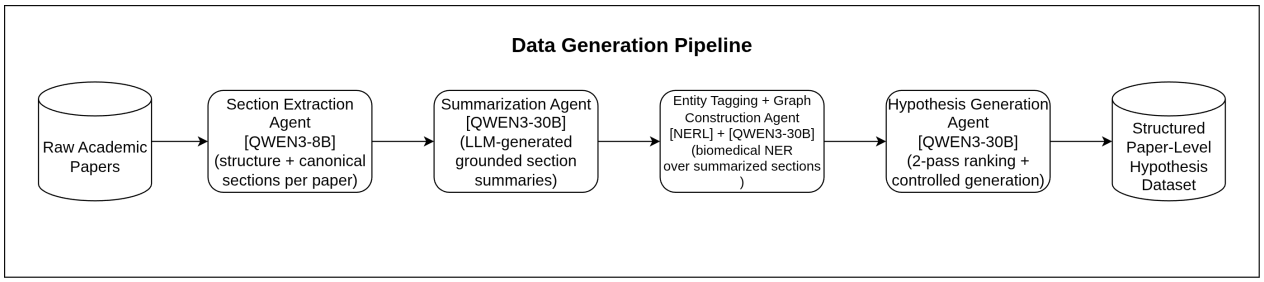

Fig. 1. Overview of the multi-agent data processing pipeline. Raw biomedical articles are sequentially processed by agents responsible for section extraction, LLM-based summarization, biomedical entity tagging, graph creation and final hypothesis generation, yielding a structured and machine-actionable dataset

combines (1) a *ranking signal* derived from the agent's salience-estimation step, and (2) a *generation signal* corresponding to the final structured hypothesis. This combined supervision encourages the model not only to produce coherent mechanistic hypotheses but also to identify which entity configurations, evidence fragments, or section-level cues provide the strongest support for those hypotheses.

Training Configuration All training runs are executed on the Qwen3 models (Qwen3 0.6B, Qwen3 1.7B, Qwen3 4B, and Qwen3 8B) [27] using parameter-efficient fine-tuning (LoRA) with bfloat16 precision and gradient checkpointing to fit multiple experiments on a single multi-GPU node. For pre-training, we use larger batch sizes and longer sequences to maximize domain exposure, whereas fine-tuning runs employ shorter sequences aligned with section-level contexts.

Hyperparameters are kept consistent across runs to allow clean comparisons of model variants. This configuration included five epochs, a learning rate of 1×10^{-4}, LoRA rank of 8 with $\alpha = 32$, bfloat16 precision, gradient checkpointing, maximum input length of 8192 tokens, and output length of 2048 tokens.

Outcome This two-stage procedure yields a compact, domain-specialized model capable of producing mechanistically grounded hypotheses from biomedical literature. The pre-training stage provides broad contextual understanding, while the supervised stage refines reasoning capabilities using structured, agent-generated supervision signals.

4.2 Validation Metrics

Validation measured hypothesis quality using BERTScore [30], which measures semantic similarity between model outputs and the ground truth. BERTScore reports precision, recall, and F1, and the results are summarized in Table 2.

4.3 Hardware Setup

Experiments were conducted on two accessible workstations (Table 1). System A was used for entity tagging and graph generation with smaller models, while System B was dedicated to large-model inference (Qwen3-32B) and LoRA training.

5 Results and Discussion

This section presents the outcomes of our experiments, focusing on how model scale and domain-specific supervision influence hypothesis generation quality. We report results across all Qwen3 variants and compare the gains introduced by supervised fine-tuning, highlighting trends that emerge consistently across model sizes.

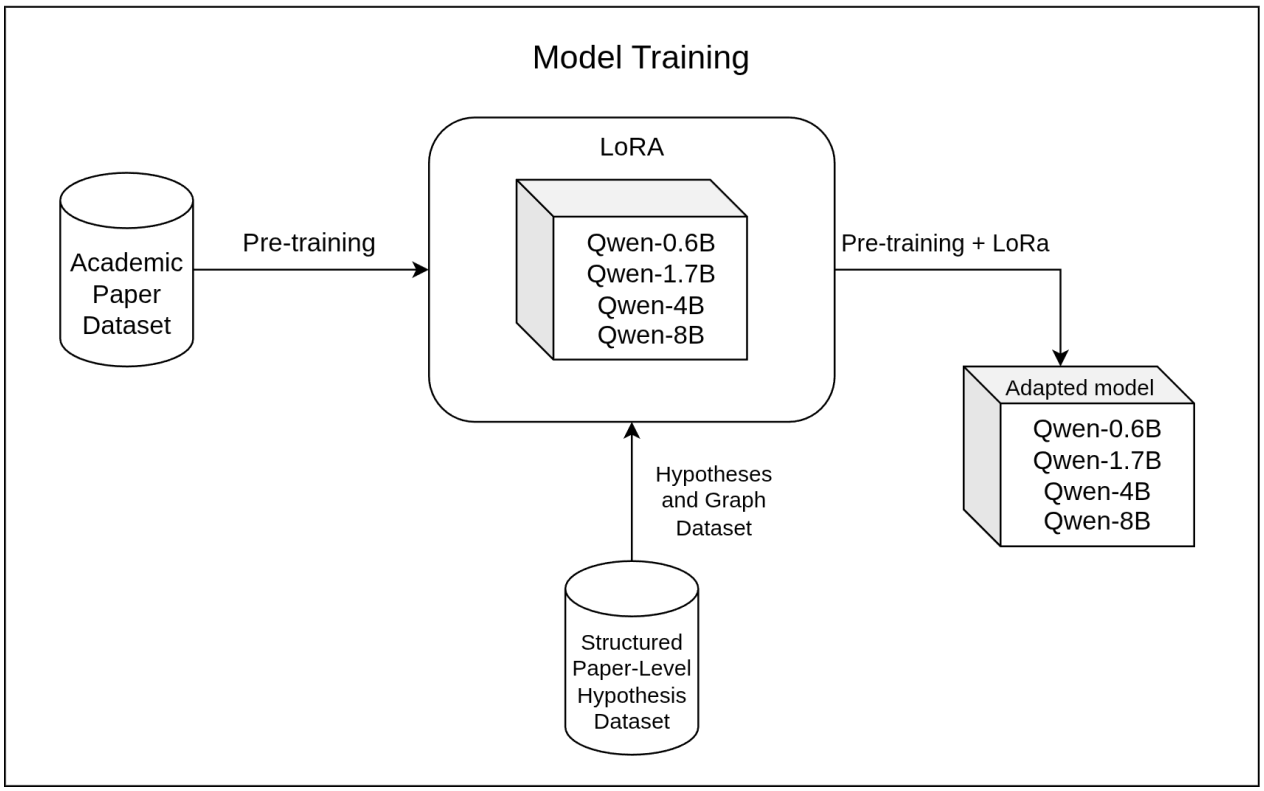

Fig. 2. Overview of the multi-stage training pipeline. The LLM first undergoes continued pre-training on full, unprocessed biomedical articles to acquire domain context and scientific language patterns. It is then fine-tuned using structured summaries, entities, graphs, and hypotheses generated by the agentic pipeline, yielding a specialized model for mechanistic hypothesis generation

5.1 Evaluation Results

We performed evaluation of all models using BERTScore [30] and ROUGE-L to capture both semantic similarity and surface-level structural alignment between generated hypotheses and reference hypotheses. The results are summarized in Table 2. Across all model sizes, supervised fine-tuning (SFT) yields substantial improvements in both semantic and lexical metrics, demonstrating the effectiveness of domain adaptation for hypothesis generation.

5.2 Interpretation of Results

The results highlight two key findings. First, the 8B model delivers the strongest overall performance, achieving the highest BERTScore F1 (91.50) and ROUGE-L (44.87). Positioning the fine-tuned 8B variant model within striking distance of a much larger 32B general model evaluated in prior work. Second, the 4B model remains surprisingly competitive: with 90.95 BERTScore F1, it trails the 8B variant by less than one point, demonstrating that mid-sized architectures can approach large-model performance once aligned with task-specific supervision.

The precision/recall trade-off further reinforces this pattern. The 8B model benefits from its larger capacity, capturing a slightly broader range of mechanistic cues and producing hypotheses with higher structural fidelity. The 4B model, while smaller, maintains strong semantic alignment and offers a balanced, efficient alternative. This continues with the trends in LLM adaptation, when

Table 1. System configurations for experimentation

System	CPU	GPU	Memory
System A	7955 WX (32) @ 5.3 GHz	1× A6000 48GB	128 GB DDR5 4800 MHz
System B	7965 WX (48) @ 5.3 GHz	2× A6000 Ada 48GB	128 GB DDR5 4800 MHz

Table 2. Evaluation results for base and fine-tuned Qwen3 models using BERTScore and ROUGE-L

Model	BERTScore P	BERTScore R	BERTScore F1	ROUGE-L F1
Qwen3-0.6B Base	86.21	89.04	87.59	26.34
Qwen3-0.6B SFT	89.56	89.80	**89.66**	35.31
Qwen3-1.7B Base	81.85	89.12	85.32	16.57
Qwen3-1.7B SFT	90.42	89.83	**90.10**	38.35
Qwen3-4B Base	82.56	88.72	85.51	18.89
Qwen3-4B SFT	90.55	91.38	**90.95**	43.45
Qwen3-8B Base	80.88	88.44	84.47	13.76
Qwen3-8B SFT	91.01	92.03	**91.50**	44.87

domain-specific fine-tuning is applied, architectural scale becomes less dominant, allowing small to mid-sized models shorten the performance gap with substantially larger models.

5.3 Implications

These results indicate that high-quality hypothesis extraction does not require extremely large models.

The fine-tuned Qwen3-4B and Qwen3-8B models achieve performance that approaches the level of a 32B model observed in previous evaluations, highlighting the power of task-aligned domain adaptation. This is particularly impactful for research environments—such as biomedical labs, hospitals, and academic groups—where access to large-scale computing infrastructure is limited.

The 4B model offers an especially attractive balance: near-large-model performance at a fraction of the memory footprint and runtime cost.

Meanwhile, the 8B variant provides the strongest overall alignment and is the most capable option for settings where maximizing recall and structural completeness is essential. Together, these findings validate that LoRA-based fine-tuning can transform mid-sized models into highly competitive scientific reasoning systems, significantly reducing the need for large foundation models.

6 Conclusion and Future Work

This work demonstrates that domain-specific fine-tuning enables small to mid-sized language models to approach the performance of far larger architectures on mechanistic hypothesis generation. The Qwen3-4B model improves from 85.51 to 90.95 BERTScore F1 after fine-tuning, narrowing the performance gap to the 8B model and reaching performance levels previously attainable only with models in the 30B+ parameter range. The 8B model further improves to 91.50 BERTScore F1 and 44.87 ROUGE-L, matching the semantic and structural fidelity of much larger systems observed in earlier studies.

These findings highlight that scaling alone does not guarantee superior performance, without alignment, large models under-perform relative to their potential. In contrast, fine-tuned small to mid-sized models demonstrate dramatic gains, becoming viable alternatives to models 4 to 8× larger. This establishes a practical pathway for research groups—with modest computational resources to develop high-performing, domain-specialized LLMs for scientific discovery.

Future work will expand these findings by:

— **Scaling up training:** Increasing the dataset to 1,000+ papers and enriching the diversity of mechanistic hypotheses.

— **Model variety:** Evaluating larger and multimodal systems (e.g., Qwen3-235B, GPT-4, Llama 3.3-70B) to characterize generalization and scaling behavior.

— **Expanded evaluation:** Incorporating BLEU, entity-level F1, and graph-based consistency metrics for a broader assessment.

— **Training strategies:** Exploring alternative PEFT methods and multi-stage tuning to further close the remaining gap with 32B+ architectures.

Acknowledgments

This research was supported by Tijuana Institute of Technology / TECNM and SECIHTI under grant number CF-2023-I-555.

References

1. **Basit, A., Hussain, K., Hanif, M. A., Shafique, M. (2024).** Medaide: Leveraging large language models for on-premise medical assistance on edge devices. https://arxiv.org/abs/2403.00830.
2. **Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., Molchanov, P. (2025).** Small language models are the future of agentic ai. https://arxiv.org/abs/2506.02153.
3. **Bornmann, L., Haunschild, R., Mutz, R. (2021).** Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications, Vol. 8, No. 1, pp. 224. DOI: 10.1057/s41599-021-00903-w.
4. **Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020).** Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, pp. 25.
5. **Corradini, F., Leonesi, M., Piangerelli, M. (2025).** State of the art and future directions of small language models: A systematic review. Big Data and Cognitive Computing, Vol. 9, No. 7. DOI: 10.3390/bdcc9070189.
6. **Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023).** Qlora: Efficient finetuning of quantized llms. https://arxiv.org/abs/2106.09685.
7. **Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D. (2023).** Gptq: Accurate post-training quantization for generative pre-trained transformers. https://arxiv.org/abs/2210.17323.
8. **Garg, M., Raza, S., Rayana, S., Liu, X., Sohn, S. (2025).** The rise of small language models in healthcare: A comprehensive survey. https://arxiv.org/abs/2504.17119.
9. **Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., Pfister, T. (2023).** Distilling step-by-step: Outperforming larger language models with less training data and smaller model sizes. https://arxiv.org/abs/2305.02301.
10. **Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W. (2021).** Lora: Low-rank adaptation of large language models. CoRR, Vol. abs/2106.09685.
11. **Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., Yoon, C., Sohn, J., Park, J., Reykhart, O., Fetherston, T., Choi, D., Kwak, S. H., Chen, Q., Kang, J. (2025).** Small language models learn enhanced reasoning skills from medical textbooks. npj Digital Medicine, Vol. 8, No. 1, pp. 240. DOI: 10.1038/s41746-025-01653-8.
12. **Lee, C., Kumar, S., Vogt, K. A., Meraj, S., Vogt, A. (2024).** Advancing complex medical

communication in arabic with sporo arasum: Surpassing existing large language models. https://arxiv.org/abs/2411.13518.

13. **Liao, Z., Wei, W., Yang, M., Kuang, X., Shi, J. (2021).** Academic publication of neurodegenerative diseases from a bibliographic perspective: A comparative scientometric analysis. Frontiers in Aging Neuroscience, Vol. 13, pp. 722944. DOI: 10.3389/fnagi.2021.722944.

14. **Liu, J., Wang, T., Liu, S., Hu, X., Tong, R., Wang, L., Xu, J. (2025).** Lightweight baselines for medical abstract classification: Distilbert with cross-entropy as a strong default. https://arxiv.org/abs/2510.10025.

15. **Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., Chen, M.-H. (2024).** Dora: Weight-decomposed low-rank adaptation. https://arxiv.org/abs/2402.09353.

16. **Moreno-García, C. F., Aceves-Martins, M., Serratosa, F. (2016).** Unsupervised machine learning application to perform a systematic review and meta-analysis in medical research. Computación y Sistemas, Vol. 20, No. 1, pp. 7–17. DOI: 10.13053/CyS-20-1-2360.

17. **Nori, H., King, N., McKinney, S. M., Carignan, D., Horvitz, E. (2023).** Capabilities of gpt-4 on medical challenge problems. https://arxiv.org/abs/2303.13375.

18. **OpenAI (2024).** Gpt-5 model. https://platform.openai.com/docs/models\#gpt-5. Accessed: 2025-11-12.

19. **Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., Dean, J. (2022).** The carbon footprint of machine learning training will plateau, then shrink. Computer, Vol. 55, No. 7, pp. 18–28. DOI: 10.1109/MC.2022.3148714.

20. **Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O. (2020).** Green ai. Communications of the ACM, Vol. 63, No. 12, pp. 54–63.

21. **Shamrikova, A. A., Krasilnikova, S. V., Ignatov, G. S., Kubysheva, N., Rudas, I., Ahmad, M., Batyrshin, I. (2025).** Application of large language models to the diagnosis of respiratory diseases. Computación y Sistemas, Vol. 29, No. 3, pp. 1865–1869. DOI: 10.13053/CyS-29-3-5925.

22. **Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., Chen, Z., Cui, J., Ding, H., Dong, M., Du, A., Du, C., Du, D., Du, Y., Fan, Y., Feng, Y., Fu, K., Gao, B., Gao, H., Gao, P., Gao, T., Gu, X., Guan, L., Guo, H., Guo, J., Hu, H., Hao, X., He, T., He, W., He, W., Hong, C., Hu, Y., Hu, Z., Huang, W., Huang, Z., Huang, Z., Jiang, T., Jiang, Z., Jin, X., Kang, Y., Lai, G., Li, C., Li, F., Li, H., Li, M., Li, W., Li, Y., Li, Y., Li, Z., Li, Z., Lin, H., Lin, X., Lin, Z., Liu, C., Liu, C., Liu, H., Liu, J., Liu, J., Liu, L., Liu, S., Liu, T. Y., Liu, T., Liu, W., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Z., Lu, E., Lu, L., Ma, S., Ma, X., Ma, Y., Mao, S., Mei, J., Men, X., Miao, Y., Pan, S., Peng, Y., Qin, R., Qu, B., Shang, Z., Shi, L., Shi, S., Song, F., Su, J., Su, Z., Sun, X., Sung, F., Tang, H., Tao, J., Teng, Q., Wang, C., Wang, D., Wang, F., Wang, H., Wang, J., Wang, J., Wang, J., Wang, S., Wang, S., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wang, Z., Wang, Z., Wei, C., Wei, Q., Wu, W., Wu, X., Wu, Y., Xiao, C., Xie, X., Xiong, W., Xu, B., Xu, J., Xu, J., Xu, L. H., Xu, L., Xu, S., Xu, W., Xu, X., Xu, Y., Xu, Z., Yan, J., Yan, Y., Yang, X., Yang, Y., Yang, Z., Yang, Z., Yang, Z., Yao, H., Yao, X., Ye, W., Ye, Z., Yin, B., Yu, L., Yuan, E., Yuan, H., Yuan, M., Zhan, H., Zhang, D., Zhang, H., Zhang, W., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, H., Zhao, Y., Zheng, H., Zheng, S., Zhou, J., Zhou, X., Zhou, Z., Zhu, Z., Zhuang, W., Zu, X. (2025).** Kimi k2: Open agentic intelligence. https://arxiv.org/abs/2507.20534.

23. **Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023).** Llama: Open

and efficient foundation language models. https://arxiv.org/abs/2302.13971.

24. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017).** Attention is all you need. CoRR, Vol. abs/1706.03762.
25. **Wang, W., Li, T., Wang, Z., Yin, Y., Zhang, S., Wang, C., Hu, X., Lu, S. (2023).** Bibliometric analysis of research on neurodegenerative diseases and single-cell rna sequencing: Opportunities and challenges. iScience, Vol. 26, No. 10, pp. 107833. DOI: 10.1016/j.isci.2023.107833.
26. **Wu, Y., et al. (2023).** Pmc-llama: Building open-source language models for medicine. arXiv preprint arXiv:2304.14454.
27. **Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z. (2025).** Qwen3 technical report. https://arxiv.org/abs/2505.09388.
28. **Yao, Z., Aminabadi, R. Y., Zhang, M., Wu, X., Li, C., He, Y. (2022).** Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. https://arxiv.org/abs/2206.01861.
29. **Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., Zhao, T. (2023).** Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. https://arxiv.org/abs/2303.10512.
30. **Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. (2019).** Bertscore: Evaluating text generation with bert. https://api.semanticscholar.org/CorpusID:127986044, Vol. abs/1904.09675.

Article received on 31/10/2025; accepted on 15/12/2025.
**Corresponding author is Claudia I. Gonzalez.*