Small Data, Small Model, Superior Domain Accuracy: Fine-Tuning a 7B Language Model as a Condensed Matter Physics Expert with 64 Examples

Junsu Park *

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1691.v1

Keywords: large language models; domain-specific fine-tuning; parameter-efficient fine-tuning; QDoRA; small data learning; condensed matter physics; graphene-metal contact resistance

 doi:10.20944/preprints202603.1691.v1

Article

Small Data, Small Model, Superior Domain Accuracy: Fine-Tuning a 7B Language Model as a Condensed Matter Physics Expert with 64 Examples

Junsu Park

Department of Physics and Photon Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Korea; incinc0987@gist.ac.kr

Abstract

Large language models excel at general scientific reasoning but struggle to reproduce the precise details of individual research papers, including specific numerical values, fabrication protocols, and mechanistic arguments. We address this limitation by fine-tuning a 7-billion-parameter model on only 64 question-answer pairs from a single condensed matter physics paper using Quantized Weight-Decomposed Low-Rank Adaptation with Chain-of-Thought reasoning (QDoRA+CoT). The fine-tuned model, evaluated under closed-book conditions, achieves 73% accuracy on 30 expert-designed questions, surpassing a 235-billion-parameter model with open-book access (60%) and the same base model with open-book access (30%). Our results demonstrate that targeted fine-tuning with minimal data on a single consumer GPU can produce domain expertise superior to models with 33 times more parameters, offering a practical pathway for researchers to build personalized AI assistants for their own work.

Keywords: large language models; domain-specific fine-tuning; parameter-efficient fine-tuning; QDoRA; small data learning; condensed matter physics; graphene-metal contact resistance

Introduction

Large language models have made significant advances in scientific reasoning, from solving mathematical problems to generating hypotheses across diverse fields [1–3]. Models with hundreds of billions of parameters can engage in sophisticated discussions about a wide range of scientific topics, drawing upon vast training data [4,5]. However, a fundamental gap exists between broad scientific literacy and the precise expertise needed to serve as a reliable assistant for a researcher's own work. The nature of this gap becomes clear when one considers what domain expertise actually requires. A condensed matter physicist studying graphene-metal contact resistance needs an assistant that knows not just the general principles of contact resistance, but the specific substrate thickness, electrode composition, channel dimensions, and fabrication sequence used in their particular experiment. It must also correctly represent the quantitative results and the theoretical framework connecting those results to physical mechanisms. Current LLMs, even with hundreds of billions of parameters, fall short of this standard. When provided with a paper's key information as an explicit prompt, large models still fabricate incorrect experimental parameters, introduce mechanisms not discussed in the paper, and confidently speculate beyond what the data supports. These are not failures of reasoning but failures of knowledge fidelity—the model fills gaps in its understanding by drawing on general training knowledge rather than faithfully representing the provided material. Existing approaches do not fully resolve this problem. Retrieval-augmented generation (RAG) can supply relevant text at inference time but remains susceptible to retrieval errors and does not prevent hallucination of details outside the retrieved context [6,7]. Full fine-tuning demands large datasets and substantial computational resources, making it impractical for individual researchers [8]. Parameter-efficient fine-tuning methods have shown strong results on general NLP benchmarks

 doi:10.20944/preprints202603.1691.v1

[9,10], yet the specific scenario of encoding a single paper's detailed knowledge—where training data is inherently scarce and evaluation requires domain expertise—has received little attention. Science-focused LLMs such as Galactica [11] and SciGLM [12] pursue broad scientific competence through large-scale pretraining, but they target breadth across disciplines rather than depth within a single study. A model that can explain general principles of contact resistance is fundamentally different from one that knows the precise fabrication conditions and measurement results of a particular experiment. In this work, we propose a lightweight pipeline for transforming a small language model into a domain expert for an individual research paper. Our approach has three components: a systematic method for constructing a compact dataset of 64 question-answer pairs organized across five knowledge categories, a fine-tuning strategy based on QDoRA+CoT that balances learning capacity with memory efficiency on a single consumer GPU, and a rigorous evaluation protocol with 30 expert-crafted questions assessed by the original paper's author. We select QDoRA specifically because it addresses two challenges simultaneously. The weight decomposition from DoRA [13] provides learning quality closer to full fine-tuning than standard LoRA, which is critical when only 64 training examples are available. Meanwhile, 4-bit quantization reduces memory requirements to fit within a single 32 GB GPU. Our experiments show that the fine-tuned 7B model, tested without access to the source paper, outperforms a 235-billion-parameter model that receives the paper's content as a prompt. Analysis of error patterns reveals that each model configuration fails in qualitatively different ways, providing insight into when and why fine-tuning on minimal data is more effective than scaling model size. These findings establish domain-specific fine-tuning as a viable and accessible strategy for building personalized scientific AI assistants, requiring only a small set of question-answer pairs and a single consumer GPU.

Methodology

The source material for this study is a condensed matter physics paper on graphene-metal contact resistance.[14–18] The paper compares two device structures—top-lead contact device (TLCD) and top-electrode contact device (TECD)—with identical channel dimensions but different contact areas, using transmission line method (TLM) and four-probe measurements on Cr/Au electrodes fabricated via a metal-on-bottom process on SiO_2/Si substrates.[19] We selected this paper as the test domain because it encompasses diverse knowledge types: device geometry, fabrication procedures, quantitative measurements, and theoretical reasoning through both the Landauer and Bardeen Transfer Hamiltonian (BTH) models. From this paper, we constructed 64 question-answer pairs organized into five categories: structure and geometry, theory and mechanism, numerical and experimental results, fabrication process, and novelty and contribution. Each answer includes Chain-of-Thought (CoT) reasoning—a step-by-step derivation preceding the final response—so the model learns logical inference chains rather than isolated facts. The dataset was prepared entirely in English to avoid multilingual interference. We fine-tuned Qwen2.5-7B-Instruct using QDoRA with rank r = 64, scaling factor α = 128, and 4-bit NF4 quantization applied to all attention and feed-forward projection layers. Training ran for 80 steps with a learning rate of 5×10^{-4} and effective batch size of 8 on a single NVIDIA RTX 5090 (32 GB VRAM), completing in approximately 20 minutes. For evaluation, we designed 30 questions spanning the same five categories but distinct from the training set. Three model configurations were compared: the fine-tuned 7B model (FT 7B) under closed-book conditions, and both the base 7B model (Base 7B) and Qwen3-235B-A22B (235B) under open-book conditions with a system prompt containing the paper's key terminology, numerical results, and theoretical framework. All responses were evaluated by the domain expert (first author of the source paper) using a three-tier rubric: correct (✓), partial (△), and wrong (✗).

Results

Table 1 summarizes the overall performance of the three model configurations across all 30 evaluation questions. The fine-tuned 7B model achieves the highest accuracy at 73%, surpassing the

 doi:10.20944/preprints202603.1691.v1

235B open-book model by 13 percentage points and outperforming the base 7B open-book model by a factor of 2.4. This result is particularly notable because the fine-tuned model operates under strictly harder conditions; it receives no information about the paper at test time, whereas both comparison models receive an explicit summary of the paper's key content. The fine-tuned model also shows the lowest proportion of partial answers (17% versus 57% for Base 7B and 30% for 235B), indicating that when it answers, it tends to do so with full accuracy rather than incomplete understanding. Table 2 presents the number of correct answers by question category for each model. The fine-tuned model demonstrates consistent superiority across all five categories. The advantage is most pronounced in the fabrication category, where FT 7B correctly answers 4 out of 6 questions compared to only 2 for the 235B model and 1 for the base model. This category requires knowledge of specific procedural details—substrate preparation, metal deposition parameters, etching methods, and contamination avoidance strategies—that are difficult to infer from general scientific knowledge. The 235B model, despite having access to relevant information in its prompt, frequently substitutes its own general knowledge of semiconductor fabrication for the specific procedures described in the paper. In the theory and mechanism category, the fine-tuned model achieves 8 out of 10, outperforming the 235B model at 7 out of 10. While the 235B model occasionally provides deeper physical insight with mathematical formulations beyond the scope of the source paper, these elaborations sometimes introduce concepts absent from the original work, resulting in partial or incorrect scores under our evaluation rubric. The fine-tuned model, by contrast, stays closely aligned with the paper's own theoretical framework.

Table 1. Overall evaluation results across 30 domain-specific questions.

Model	Parameters	Condition	Correct (√)	Partial (△)	Wrong (✗)	Accuracy
Base 7B	7B	Open-book	9	17	4	30%
FT 7B	**7B**	**Closed-book**	**22**	**5**	**3**	**73%**
235B	235B	Open-book	18	9	3	60%

Table 2. Correct answers (✓) by question category.

Category	Questions	Base 7B	FT 7B	235B
Structure / Geometry	4	1	3	3
Theory / Mechanism	10	4	8	7
Numerical / Experimental	6	2	5	4
Fabrication	6	1	4	2
Novelty / Contribution	4	1	2	2

The numerical and experimental category further highlights the fine-tuned model's strength in factual precision. The fine-tuned model correctly reproduces specific values such as substrate dimensions, dielectric thickness, electrode composition, and channel lengths that the base model fails to recall and the 235B model occasionally fabricates. This pattern suggests that fine-tuning effectively encodes precise quantitative information into the model's parameters, whereas large models struggle to faithfully extract such details from their input context even when explicitly provided. Beyond

 doi:10.20944/preprints202603.1691.v1

overall accuracy, the three models exhibit qualitatively distinct error patterns that reveal different failure mechanisms. The base 7B model, despite receiving the paper's key information as a prompt, shows a dominant pattern of knowledge absence. It produces 17 partial answers out of 30, indicating that it can generate generally relevant responses but consistently lacks the specificity required for correct answers. For example, when asked about the SiO_2 dielectric thickness, the base model states that the thickness "was not specified in the given information," even though this detail was available in the paper and could have been included in a more thorough reading of the domain. Its errors reflect an inability to absorb and retain fine-grained details from the input context. The 235B model exhibits a pattern of confident over-extrapolation. When the provided prompt does not contain sufficient detail to answer a question fully, the model fills the gap with plausible scientific knowledge drawn from its pretraining data. In several instances, it introduces physical mechanisms such as Fermi-level pinning, carbon-metal interdiffusion, and Schottky barrier effects at contact perimeters—none of which are discussed in the source paper. It also fabricated incorrect experimental parameters, such as channel lengths of 1–10 μm when the actual values were 40–120 μm, and incorrectly claimed that reactive ion etching was not used in the fabrication process. These errors are particularly problematic because they are scientifically plausible, making them difficult to detect without expert knowledge of the specific paper. The fine-tuned 7B model shows a different and more targeted failure mode: directional reversal of learned relationships. In three questions marked as wrong, the model correctly identified the relevant physical quantities but reversed their relative contributions. For instance, when asked how specific contact resistivity (ϱc) and edge resistance (Rgg') contribute differently in TLCD versus TECD devices, the model attributed ϱc dominance to TLCD and Rgg' dominance to TECD—the exact opposite of the correct relationship. Notably, the same model answered a closely related question about the same physical quantities correctly in a different question context. This pattern suggests that the model has learned the relevant concepts but occasionally activates them in the wrong configuration, a hallucination mode distinct from the fabrication of entirely novel information seen in the 235B model. Such directional reversals may be addressable through targeted data augmentation that reinforces the correct relational structure between key concepts.

Discussion

The central finding of this study—that a fine-tuned 7B model under closed-book conditions outperforms a 235B model with open-book access—challenges the common assumption that larger models with more context yield better domain-specific performance. This result can be understood by examining how each model handles the boundary between what it knows and what it does not. The 235B model possesses broad scientific knowledge and generates fluent, well-structured responses. However, when the provided prompt leaves gaps in coverage, the model seamlessly fills those gaps with general knowledge from its pretraining data, producing answers that are scientifically reasonable but factually inconsistent with the source paper. The model has no mechanism to distinguish between information drawn from the prompt and information drawn from its own parameters, leading to confident over-extrapolation that is difficult to detect without expert verification. The fine-tuned model, by contrast, has internalized the paper's specific content through parameter updates. Its knowledge, while narrower, is directly encoded rather than retrieved from a transient context window, resulting in higher fidelity for the target domain. This interpretation also explains the distinct error patterns observed across models. The base 7B model lacks the capacity to fully absorb complex domain information from a single prompt, resulting in vague and incomplete responses. The 235B model has sufficient capacity but conflates its vast general knowledge with the specific details it was given, producing a mixture of correct recall and plausible fabrication. The fine-tuned model avoids both of these failure modes for most questions, but occasionally reverses the directionality of learned relationships—a more subtle form of hallucination that likely arises from insufficient training signal to disambiguate closely related concepts. This suggests that the 64-example training set, while remarkably effective overall, reaches its limits at the boundaries of

 doi:10.20944/preprints202603.1691.v1

conceptual similarity where relational structure must be precisely encoded. From a practical standpoint, our pipeline requires minimal resources. A researcher can construct a dataset of 64 question-answer pairs in two to three hours, and training completes in approximately 20 minutes on a single consumer GPU. The resulting model can then serve as a personalized assistant for dissertation defense preparation, student training, peer review responses, or long-term knowledge preservation for a research group. Unlike RAG-based systems, the fine-tuned model requires no retrieval infrastructure at inference time and carries no risk of retrieval failure. Several limitations should be noted. This study validates the approach on a single paper from one subdomain of condensed matter physics, and generalization to other scientific fields remains to be demonstrated. The evaluation relies on manual expert assessment, which, while rigorous, does not scale easily. The fine-tuned model's directional reversal errors indicate that hallucination is reduced but not eliminated. Future work should explore multi-paper fine-tuning, automated evaluation using LLM-as-judge frameworks, and extension to other scientific domains to establish the generality of these findings.

Conclusion

We have demonstrated that fine-tuning a 7-billion-parameter model on 64 question-answer pairs from a single research paper produces domain expertise that surpasses a 33-times-larger model with explicit access to the source material. The fine-tuned model achieves 73% accuracy under closed-book conditions, compared to 60% for the 235B open-book model and 30% for the base 7B open-book model. These results establish that targeted fine-tuning with minimal data on a single consumer GPU is a viable strategy for building personalized scientific AI assistants, lowering the barrier for researchers to create domain-specific models tailored to their own work.

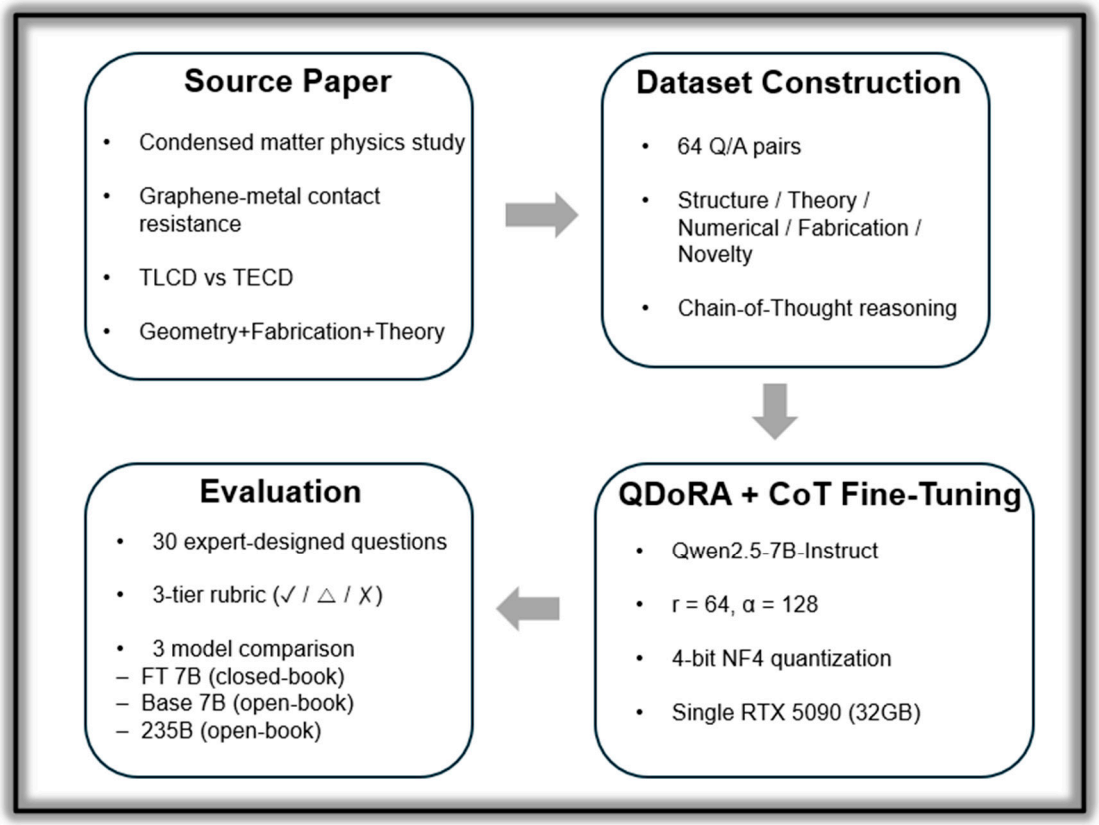


Figure 1. Overview of the proposed pipeline for building a domain-specific language model. A condensed matter physics paper on graphene–metal contact resistance serves as the source material for constructing a compact dataset of 64 question–answer pairs across five categories (structure, theory, numerical results, fabrication, and

 doi:10.20944/preprints202603.1691.v1

novelty). The dataset is used to fine-tune a Qwen2.5-7B-Instruct model using QDoRA with Chain-of-Thought reasoning under 4-bit NF4 quantization on a single RTX 5090 GPU. Model performance is then evaluated using 30 expert-designed questions under a three-tier rubric (✓ correct, △ partial, ✗ wrong) by comparing the fine-tuned 7B model with the base 7B and a 235B model under open-book conditions.

a)

	Question	Base 7B (Open-book)	FT 7B (Closed-book)	Qwen3-235B (Open-book)
Structure / Geometry				
Q1	TLCD/TECD structure	✓	✓	✓
Theory / Mechanism				
Q2	Why larger area reduces R_c	△	✓	✓
Q4	BTH model necessity	✓	✓	✓
Q5	Sub-linear relationship	△	△	△
Q6	ρ_c vs $R_{gg'}$ distinction	△	✗	✓
Q9	Edge crowding insufficient	△	✓	△
Numerical / Experimental				
Q3	4-probe values (487/125 Ω)	✓	✓	✓
Q7	Gate voltage modulation	△	△	✓
Fabrication				
Q8	Metal-on-bottom process	△	✓	△
Novelty / Contribution				
Q10	Core novelty of comparison	✓	✓	✓
⋮	⋮	⋮	⋮	⋮

✓ Correct △ Partial ✗ Wrong

b)

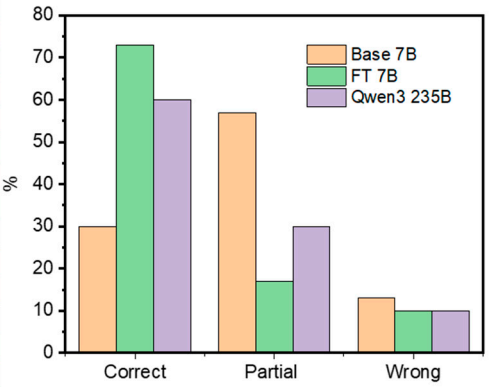
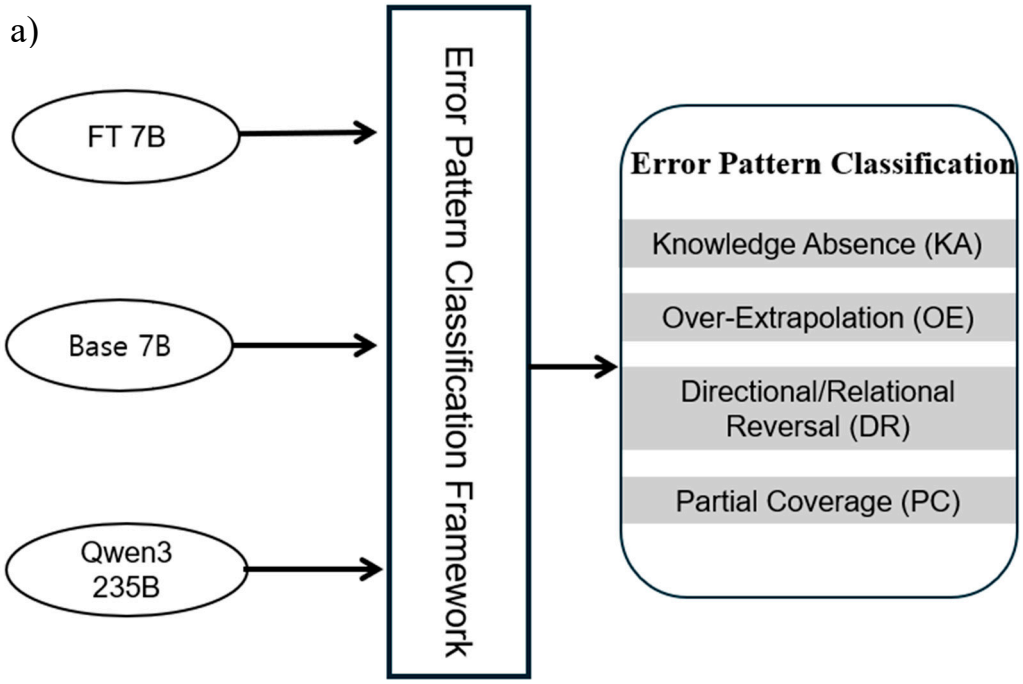


Figure 2. Evaluation results comparing the fine-tuned 7B model with baseline models. **(a)** Representative question-level assessment across different knowledge categories, including structure/geometry, theory/mechanism, numerical/experimental results, fabrication, and novelty. Each response is evaluated using a three-tier rubric: correct (✓), partial (△), and wrong (✗). The comparison includes the base 7B model with open-book prompting, the fine-tuned 7B model evaluated under closed-book conditions, and the 235B model with open-book access. **(b)** Overall distribution of evaluation outcomes across all questions. The fine-tuned 7B model shows the highest proportion of correct answers and the lowest proportion of partial responses, demonstrating improved fidelity to the source paper compared with both the base 7B model and the larger 235B model.



 doi:10.20944/preprints202603.1691.v1

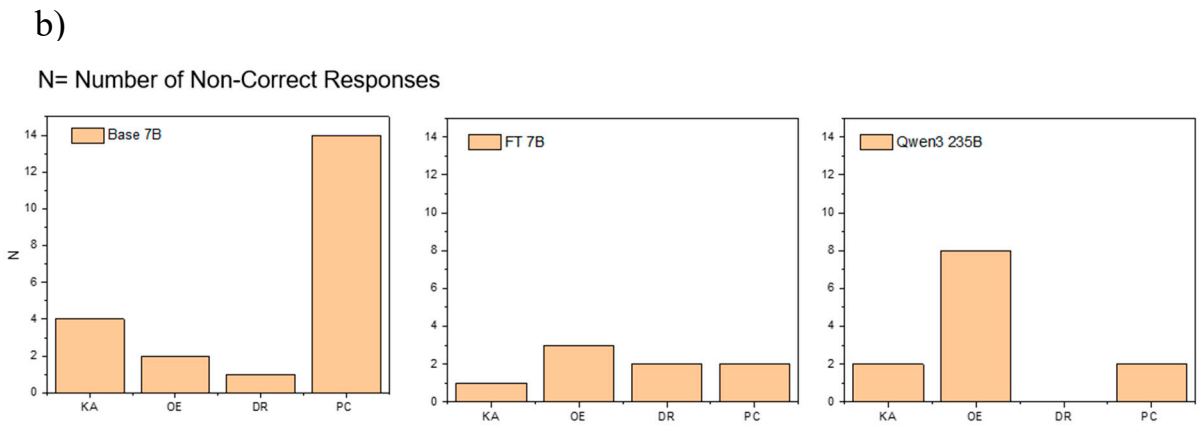


Figure 3. Error pattern analysis of the evaluated language models. **(a)** Schematic illustration of the error classification framework used to analyze non-correct responses from the fine-tuned 7B model (FT 7B), the base 7B model, and the 235B model. Errors are categorized into four types: knowledge absence (KA), where the model lacks the necessary domain information; over-extrapolation (OE), where the model introduces plausible but unsupported scientific explanations; directional/relational reversal (DR), where the relevant concepts are recalled but their relationships are inverted; and partial coverage (PC), where the response captures general ideas but omits critical details. **(b)** Distribution of non-correct responses (N) across the four error categories for each model. The base 7B model is dominated by partial coverage errors, the 235B model frequently exhibits over-extrapolation, and the fine-tuned 7B model shows fewer overall errors with occasional relational reversals. These distinct patterns reveal different failure mechanisms and illustrate how targeted fine-tuning improves domain-specific knowledge fidelity.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Acknowledgments: This work was conducted independently by the author using a personal workstation and AI-assisted tools for dataset construction and model development. No external funding supported this study.

AI Usage Statement: AI-assisted tools were used for code generation, dataset preparation, and language editing under the supervision of the author.

Data and Code Availability: The datasets, training scripts, and evaluation code used in this study are publicly available in the GitHub repository at https://github.com/parkjeff27-art/graphene-contact-LLM. The final fine-tuned model weights (QDoRA+CoT configuration) have been uploaded to Hugging Face and can be accessed at https://huggingface.co/pjspjs0987/Qwen2.5-7B-GrapheneContact-QDoRA-CoT. Detailed inference transcripts and the evaluation rubric are provided in the Supplementary Material.

Reference

1. Zhang, Y., Khan, S. A., Mahmud, A., Yang, H., Lavin, A., Levin, M., Frey, J., Dunnmon, J., Evans, J., Bundy, A., Dzeroski, S., Tegner, J. & Zenil, H. Exploring the role of large language models in the scientific method: from hypothesis to discovery. npj Artif. Intell. 1, 14 (2025).
2. Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., Schulz, E. How should the advancement of large language models affect the practice of science? Proc. Natl Acad. Sci. U S A 122, e2401227121 (2025).
3. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. Nat. Med. 29, 2983–2984 (2023).
4. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. Proc. Natl Acad. Sci. U S A 120, e2215907120 (2023).

 doi:10.20944/preprints202603.1691.v1

5. Xiao, C., Cai, J., Zhao, W., Lin, B., Zeng, G., Zhou, J., Zheng, Z., Han, X., Liu, Z., Sun, M. Densing law of LLMs. Nat. Mach. Intell. 7, 1823–1833 (2025).
6. Omar, M., Sorin, V., Collins, J. D., Reich, D., Freeman, R., Gavin, N., Charney, A., Stump, L., Bragazzi, N. L., Nadkarni, G. N. & Klang, E. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun. Med. 5, 330 (2025).
7. Yang, R., Ning, Y., Keppo, E., Liu, M., Hong, C., Bitterman, D. S., Ong, J. C. L., Ting, D. S. W. & Liu, N. Retrieval-augmented generation for generative artificial intelligence in health care. npj Health Syst. 2, 2 (2025).
8. Prottasha, N. J., Mahmud, A., Sobuj, M. S. I., Bhat, P., Kowsher, M., Yousefi, N. & Garibay, O. O. Parameter-efficient fine-tuning of large language models using semantic knowledge tuning. Sci. Rep. 14, 30667 (2024).
9. Qin, Y., Su, Y., Hu, S., Chen, W., Yi, J., Zhao, W., Tang, J., Ding, N., Yang, G., Wei, F., Yang, Z., Chen, Y., Chan, C.-M., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Li, J. & Sun, M. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat. Mach. Intell. 5, 220–235 (2023).
10. Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S. & Yang, F. Parameter-efficient fine-tuning in large language models: a survey of methodologies. Artif. Intell. Rev. 58, 227 (2025).
11. Chartier-Edwards, N., Grenier, E. & Goujon, V. Galactica's dis-assemblage: Meta's beta and the omega of post-human science. AI Soc. 40, 2069–2081 (2025).
12. Asai, A., He, J., Shao, R., Shi, W., Singh, A., Chang, J. C., Lo, K., Soldaini, L., Feldman, S., D'Arcy, M. et al. Synthesizing scientific literature with retrieval-augmented language models. Nature 650, 857–863 (2026).
13. Zhu, H., Yang, H., Wang, Y., Hu, K., He, G., Zhou, J., Li, Z. & Alzheimer's Disease Neuroimaging Initiative. A new low-rank adaptation method for brain structure and metastasis segmentation via decoupled principal weight direction and magnitude. Sci. Rep. 15, 27388 (2025).
14. W. Liu, J. Wel, X. Sun and H. Hongyu, A Study on Graphene—Metal Contact, Crystals 3, 257 (2013)
15. T. Cusati, G. Fiori, A. Gahoi, V. Passi, M. C. Lemme and G. Iannaccone, Electrical Properties of Graphene-Metal Contacts, Sci. Rep. 7, 5109 (2017).
16. A. Gahoi, S. Kataria, Andreas Bablich, Satender Kataria, Vikram Passi, M. C. Lemme, Contact Resistance Study of Various Metal Electrodes with CVD Graphene, Solid-State Electronics 125, 4063 (2016).
17. S. Russo, M. F. Craciun, M. Yamamoto, A. F. Morpurgo and S. Tarucha, Contact Resistance in Graphene-Based Devices, Phys. E 42, 677 (2010).
18. F. Urban, G. Lupina, A. Grillo, N. Martucciello and A. Di Bartolomeo, Contact Resistance and Mobility in Back-Gate Graphene Transistors, Nano Express 1, 010001 (2020).
19. Park, J., Lee, J.S. & Lee, S. Beyond edge injection: area effects in graphene–metal contact resistance. J. Korean Phys. Soc. (2026).