



Benchmarking and fine-tuning large language models for domain-specific materials science: a case study toward GPU application

Yuqin Wang^{1,2}, Youle Wang², Lei Zhang^{1,2}

Keywords:

Large language model, fine-tuning, GPU materials, domain LLM, benchmark, corpus

Citation:

Wang, Y.; Wang, Y.; Zhang, L. Benchmarking and fine-tuning large language models for domain-specific materials science: a case study toward GPU application. *AI Agent* 2026, 2, 26. <https://dx.doi.org/10.20517/aiagent.2026.43>

Received: 13 Jul 2026

First Decision: 7 Aug 2026

Revised: 20 Aug 2026

Accepted: 28 Sep 2026

Published: 26 Oct 2026

Academic Editor:

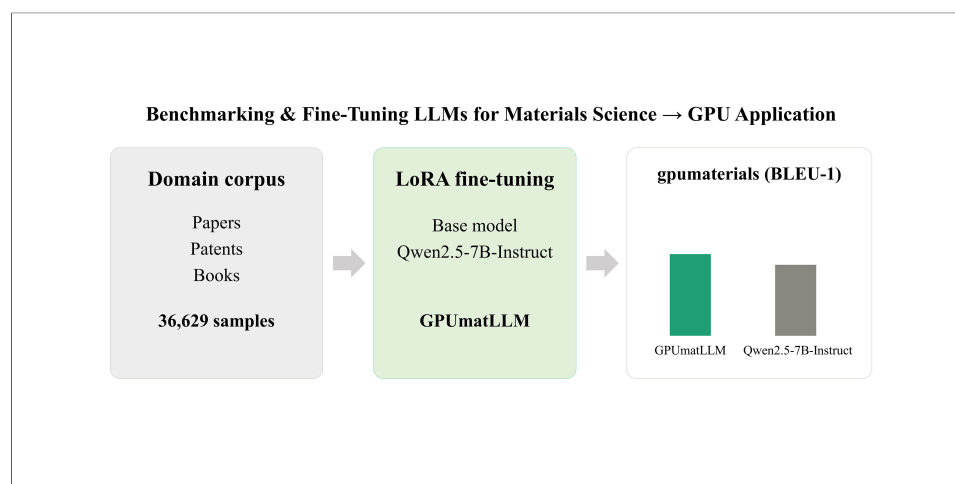
Jianjun Liu

Copy Editor:

Tong Wang

Production Editor:

Tong Wang



Abstract

Despite significant progress in large language models (LLMs) with general reasoning capabilities, their adaptability remains limited in interdisciplinary domains that require specialized expertise, such as those related to cutting-edge chip fabrication in materials science. To address this gap, this study constructs a domain-specific benchmark for graphics processing unit (GPU) materials science, comprising a systematically constructed evaluation dataset and a comprehensive assessment of representative LLMs on domain-specific tasks. As part of this benchmark development, we also develop GPUmatLLM, a specialized domain language model fine-tuned on high-quality data from academic papers, patents, and other relevant sources for GPU materials science. Benchmarking results reveal a notable performance gap among the representative LLMs, with no single model excelling in both general scientific reasoning and GPU material-specific knowledge acquisition. In contrast, GPUmatLLM demonstrates improved comprehensive performance on GPU materials tasks, surpassing baseline models in domain-specific inference while maintaining competitive performance in general scientific reasoning. This work provides an effective framework and practical paradigm for benchmarking and optimizing domain-specific LLMs in specialized materials science fields.

¹Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics of the Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China.

²School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, Jiangsu, China.

Correspondence to: Dr. Lei Zhang, Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics of the Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China. E-mail: zhanglei1@opt.ac.cn

INTRODUCTION

The rapid advancement of large language models (LLMs) has positioned artificial intelligence at the frontier of scientific knowledge management^[1,2], offering unprecedented opportunities to automate the retrieval, synthesis, and reasoning of technical literature across diverse disciplines^[3,4]. LLMs are increasingly regarded as scientific reasoning engines rather than static knowledge stores^[5,6]. Comparable artificial intelligence (AI)-driven advances are reshaping adjacent hardware domains, including optical image processing^[7], metasurface design^[8] and fiber-optic signal processing^[9]. These AI systems are built on the Transformer architecture^[10] and, in their most recent form, on large open-weight models^[11]. Their deployment at scale has raised questions of reliability and risk control in the physical sciences^[12]. The performance, reliability, and scalability of this hardware ultimately depend on materials, which has sparked a wave of research in graphics processing unit (GPU) manufacturing processes. In the field of GPU materials science [e.g., encompassing thermal interface materials (TIMs), semiconductor packaging, heat dissipation structures, and manufacturing processes], research output has expanded dramatically in recent years, driven by escalating thermal challenges in high-power GPU devices and the proliferation of candidate material systems^[13,14]. This domain is inherently interdisciplinary^[15,16], integrating polymer chemistry, phonon transport physics, semiconductor packaging engineering, and manufacturing reliability, with practical knowledge scattered across academic journals^[17], patent databases^[18], and industrial standards^[19]. Efficiently retrieving and reasoning over such fragmented, cross-domain knowledge is critical for accelerating material selection^[20,21], process optimization^[22,23], and failure analysis in GPU thermal management applications^[24,25].

Traditional knowledge acquisition in materials science remains predominantly manual^[26,27], relying on keyword-based database searches^[28], labor-intensive literature reviews^[29,30], and expert consultation^[31]. These workflows are slow, difficult to scale^[32,33], and prone to coverage gaps as the body of literature continues to grow^[34]. While general-purpose LLMs have shown promise in scientific information extraction^[35,36] and reasoning^[37], their performance on domain-specific materials queries, such as those arising in GPU manufacturing, remains unreliable^[38,39]. Without targeted exposure to the specialized terminology, physical constraints, and engineering trade-offs of GPU thermal management, general LLMs often generate superficial, generic, or factually inaccurate responses^[40,41], limiting their utility for high-stakes engineering decision-making^[42]. Domain-adapted materials chemistry models, such as ChemLLM and ChemDFM, represent early efforts toward scientific specialization, but their training distributions are centered on molecular chemistry and pharmaceutical science rather than the device-level packaging and thermal transport challenges central to GPU materials engineering. Beyond single-model approaches, recent work has begun to integrate curated materials databases, large-scale literature mining and agent-based workflows into unified discovery pipelines^[43], and to couple domain-specific models with autonomous experimental planning. These developments underline the importance of a well-characterized domain corpus and of transparent, domain-matched benchmarks, which are the focus of the present work.

Despite growing interest in applying LLMs to materials science, several critical barriers, which remain underexplored in the existing literature, hinder the development of a robust GPU materials-specific LLM^[4,44]. First, this highly interdisciplinary field lacks a systematically curated, large-scale training corpus; constructing such a corpus requires automated parsing, quality control, and question-answer (QA) generation pipelines to integrate heterogeneous document types^[45,46]. These include peer-reviewed papers, technical books, and patents; curated collections of such heterogeneous sources for GPU materials engineering are currently lacking^[47,48]. Second, pure domain-specific fine-tuning risks catastrophic forgetting

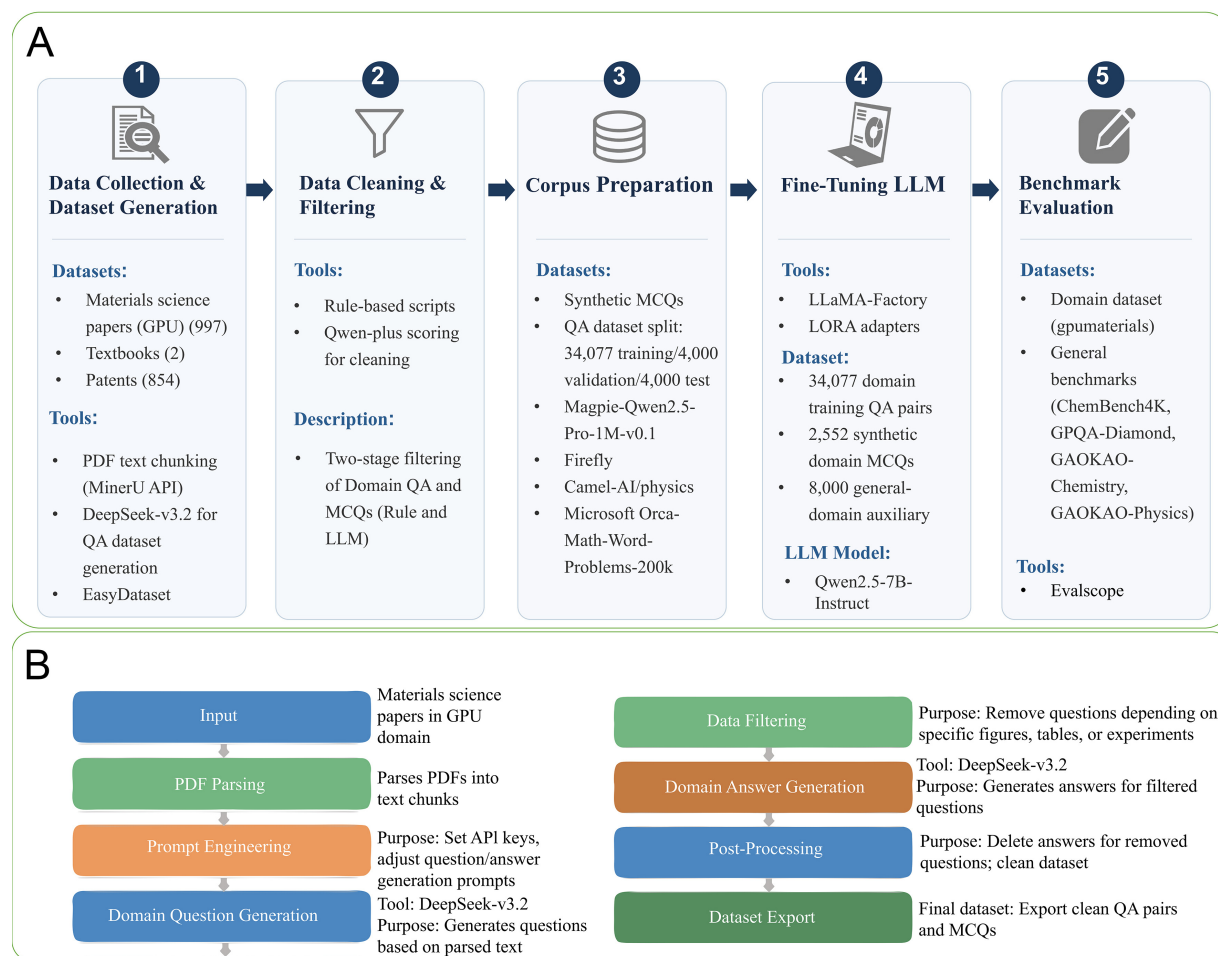


Figure 1. (A) Overall five-stage workflow of GPUmatLLM, covering data collection and dataset generation, data cleaning and filtering, corpus preparation, model fine-tuning, and benchmark evaluation; (B) Automated QA and MCQ dataset generation pipeline for GPUmatLLM. GPU: Graphics processing unit; API: application programming interface; QA: question-answer; MCQs: multiple-choice questions; LLM: large language model; GPQA: graduate-level Google-proof Q&A; LoRA: low-rank adaptation; PDF: portable document format.

of general language and scientific reasoning capabilities^[49,50], yet optimal data mixing strategies for narrow engineering domains with limited corpus size remain understudied^[51]. Third, standardized LLM benchmarks for GPU materials are absent, impeding objective evaluation and transparent comparison of model performance across different architectures and training paradigms^[52,53].

To address these gaps, this work introduces GPUmatLLM, with a domain-specific benchmarking dataset and an LLM tailored for GPU materials science, built upon curated corpus construction, parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA), and systematic benchmark evaluation. A training corpus of 44,629 samples, comprising 34,077 domain QA pairs derived from 997 peer-reviewed articles, 2 technical books, and 854 GPU-related patents, together with 2,552 synthesized domain multiple-choice questions (MCQs) and 8,000 general-domain auxiliary samples, is assembled through an automated filtering pipeline. The Qwen2.5-7B-Instruct base model is fine-tuned to generate GPUmatLLM, which is evaluated on four standardized benchmarks against six representative models. The benchmarking spans general, chemistry-specialized, and state-of-the-art LLMs, using EvalScope as a unified evaluation framework. The overall workflow, covering corpus construction, LoRA fine-tuning and benchmark evaluation, is summarised in Figure 1. A preliminary version of this work has been made available as a preprint^[54]. This work represents a systematic attempt to develop a dedicated benchmark and LLM for GPU-related materials science.

METHODS

Construction of the GPU materials corpus and benchmark

A domain corpus for the GPU materials field was assembled from three complementary sources. First, 997 peer-reviewed papers were retrieved from the OpenAlex database using core search topics covering TIMs, GPU thermal management, semiconductor packaging, die-attach processes, and materials characterization. Then, 854 GPU-related patents were collected from Google Patents, and two books on electronic thermal management were incorporated to supply systematic disciplinary knowledge^[55,56].

To convert these heterogeneous documents into usable text, every raw portable document format (PDF) file was parsed into structured Markdown through the MinerU application programming interface (API) integrated within the EasyDataset platform. During parsing, structure-aware semantic chunking was applied instead of fixed-length segmentation, so that quantitative parameters spanning several sentences (e.g., thermal conductivity, bond-line thickness, and process-temperature ranges) remained together with their surrounding context.

To prevent the model from losing its general capabilities during subsequent domain adaptation, 8,000 non-domain auxiliary samples were introduced: general-instruction data from the Magpie^[57] and Firefly Chinese corpora, 3,000 physics-reasoning samples from Camel-AI/physics^[58], and 3,000 mathematical word problems from Orca-Math^[59].

From each resulting text chunk, single-turn QA pairs and MCQs were generated by prompting DeepSeek-V3.2 through the SiliconFlow API. A separate prompt template was designed for each data type (i.e., QA pairs from papers, QA pairs from patents, and MCQs) with the full templates provided in the Supporting Information. After the two-stage filtering pipeline, 42,077 QA pairs and 2,552 MCQs were retained. The prompt templates used for QA and MCQ generation are provided in Prompts 1-4 in [Supplementary Materials](#). The raw samples were then refined through a two-stage hierarchical filtering pipeline [[Supplementary Figure 1](#)]^[60]. In the first stage, rule-based filtering within the EasyDataset platform^[61] was used to remove items built on non-knowledge fragments, including figure captions, table headers, author information, and bibliographic references. In the second stage, the Qwen-plus model scored each remaining sample from 1 to 5 on factual correctness, information density, and logical coherence; samples scoring below 12 out of 15 in total, or below 4 for factual correctness, were discarded. No independent manual audit of the retained samples was performed; the quality characteristics reported here are therefore those of the automated scorer rather than of human adjudication. This is a deliberate consequence of the fully automated design and is noted among the limitations of the present study.

After filtering, 42,077 high-quality domain QA pairs remained. Because each text chunk yielded several QA pairs, the 42,077 filtered pairs were partitioned into training (34,077), validation (4,000), and test (4,000) subsets by stratified sampling within each chunk. This internal partition is chunk-overlapping by construction, since QA pairs derived from the same chunk may fall into different subsets. It was used solely to monitor training convergence and to select the final checkpoint, and no result reported in this work is computed on it. All reported domain results are obtained on the gpumaterials benchmark, which was held out and never used for fine-tuning. Combining the training split with the auxiliary data (i.e., 8,000 samples) and the synthesized MCQs (i.e., 2,552 samples, *vide infra*) yielded the final training corpus of 44,629 samples. Independently, a held-out evaluation benchmark, gpumaterials, was constructed, consisting of 500 open-ended QA pairs with model-generated reference answers spanning TIMs, semiconductor substrates, interconnect metallurgy, and packaging polymers. Disjointness between the benchmark and the training corpus was enforced at the item level and was quantified as follows. For every benchmark item, the maximum character n-gram (3-5) term frequency-inverse document frequency (TF-IDF) cosine similarity

against all 34,077 training samples was computed. The distribution is concentrated at low values, with a median of 0.289, a mean of 0.305, a 95th percentile of 0.552, and a 99th percentile of 0.671. A single item (0.2% of the benchmark) exceeds a similarity of 0.80, reaching 0.854, in which the same measurement had been reformulated as two separate questions; because it constitutes only 0.2% of the benchmark, it cannot materially affect any reported mean or model ranking. This analysis provides an upper bound on lexical overlap; it would not detect a heavily paraphrased restatement of the same fact, and document-level provenance was not retained during corpus construction, so source-file-level disjointness cannot be verified retrospectively.

The gpumaterials benchmark was constructed using the same automated pipeline as the training corpus rather than authored by hand. Candidate QA pairs were generated by DeepSeek-V3.2 from GPU materials source chunks using the prompt template specified in Prompt 3 in [Supplementary Materials](#), and then filtered in two stages: rule-based removal of non-knowledge fragments, followed by Qwen-plus scoring under the same acceptance thresholds described above. Items that could not be answered from text alone were removed at the rule-based stage. Because the generator and the scorer belong to different model families, each retained item is validated independently of the model that produced it, and each item is paired with a single reference answer. Reference answers are written in Chinese and have a median length of 423 characters (interquartile range 293–593, full range 18–2007). Because bilingual evaluation understudy, 1-gram (BLEU-1) on Chinese text depends on the segmentation used, we report the tokenization adopted by the evaluation framework in [Supplementary Table 1](#). Since BLEU-1 is computed against a single reference and no equivalence matching of paraphrased wording, units, or numerical formats is applied, it is reported here as a lexical-fidelity proxy rather than a measure of semantic correctness; scores nevertheless remain comparable across models, which are evaluated against identical references under identical decoding settings.

Domain-specific fine-tuning

Qwen2.5-7B-Instruct was adopted as the base model and adapted to the GPU materials domain through LoRA within the LLaMA-Factory framework^[62]. LoRA^[63] expresses the weight update ΔW as the product of two low-rank matrices, as defined in:

$$\Delta W = BA \quad (1)$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$. The pre-trained weights W_o were kept frozen while only A and B were trained, so that the forward computation became:

$$h = W_{0x} + BAx \quad (2)$$

where d and k denote the output and input dimensions of the pre-trained weight matrix W_o , x is the layer input, and h is the layer output, thereby preserving the general representations of the base model at the parameter level. The complete fine-tuning workflow is illustrated in [Supplementary Figure 2](#).

LoRA adapters of rank $r = 8$ were attached to the query and value projection layers of every Transformer block. Training used a learning rate of 1×10^{-4} under a cosine-decay schedule, a batch size of 4 with 4-step gradient accumulation (an effective batch size of 16), brain floating point 16 (BF16) mixed precision, and a single epoch over the corpus. All fine-tuning was carried out on a single NVIDIA GeForce RTX 5090 GPU with 32 GB of memory.

Table 1. Baseline models and selection rationale.

Model	Category	Descriptions
Qwen2.5-7B-Instruct	General	Base model; quantifies the absolute performance gain of LoRA Fine-Tuning ^[64]
Llama-3.1-8B-Instruct	General	General-purpose baseline enabling cross-study comparison at similar parameter scale ^[65]
ChemDFM-v1.5-8B	Specialized	Chemistry/materials domain expert with third-party evaluation data ^[66]
ChemLLM-7B-Chat	Specialized	Chemistry/materials fine-tuned model; a highly relevant Domain-Specific competitor ^[67]
Qwen3-32B	Frontier	Leading 30B-scale Chinese language model; establishes large model performance ceiling ^[68]
DeepSeek-V3.2	Frontier	State-of-the-art frontier model; serves as the upper-bound reference ^[69]

Table 2. Benchmark datasets used for model evaluation.

Dataset	Descriptions
ChemBench4K	A chemistry LLM benchmark comprising 4,100 High-quality single-choice questions across nine core chemistry task categories ^[67]
GAOKAO_Bench (Chemistry & Physics MCQ)	A benchmark based on Chinese National College Entrance Examination (Gaokao) problems, evaluating language comprehension and logical reasoning ^[70] . Only the multiple-choice chemistry and physics subsets were used, as they better reflect domain scientific reasoning
GPQA-Diamond	A dataset of 198 expert-written MCQs in biology, physics, and chemistry, designed to challenge graduate-level domain understanding ^[71]
gpmaterials	500 open-ended QA pairs on GPU materials and processing, constructed in this work. Evaluates the direct effect of domain-specific fine-tuning

LLM: Large language model; MCQs: multiple-choice questions.

Benchmark evaluation protocol

GPUmatLLM was evaluated against six baseline models spanning three categories, general-purpose LLMs, chemistry-specialized LLMs, and large-scale frontier models; the complete list and the role of each baseline are reported in Table 1. To ensure a fair and reproducible comparison, every model was run within the unified EvalScope framework, which standardizes prompt construction, response parsing, and metric computation across models. The decoding strategy was fixed to greedy decoding with the temperature set to 0 for all models. The lightweight 7-8B models were run locally on the RTX 5090 under BF16 precision, while Qwen3-32B was accessed through Alibaba Cloud and DeepSeek-V3.2 through SiliconFlow, as their parameter scales exceeded the local memory capacity.

Every model was then assessed on the four benchmark datasets summarized in Table 2. Because GAOKAO_Bench contributes separate chemistry and physics subsets, these four datasets yield five evaluation sets in total: four multiple-choice sets scored by accuracy (GPQA-Diamond, GAOKAO-Chemistry, GAOKAO-Physics, ChemBench4K) and one open-ended set scored by BLEU-1 (gpmaterials).

ChemBench4K and GAOKAO_Bench were downloaded and converted into the MCQ format recognized by EvalScope, while GPQA-Diamond is natively provided within EvalScope. The first three assess general scientific and chemical reasoning, whereas gpmaterials measures open-ended generation in the target domain. For the MCQ benchmarks, accuracy is reported, defined as the proportion of questions whose selected option matches the ground-truth answer. For gpmaterials, where each question is paired with a reference answer, BLEU-1 is adopted as the evaluation metric. The overall procedure is summarized in Figure 2. Beyond these quantitative scores, a qualitative comparison of representative responses was carried out to confirm that the measured differences reflect genuine domain reasoning rather than scoring artifacts. Complete model versions, hyperparameters, decoding settings, and API access dates are listed in Supplementary Table 1.

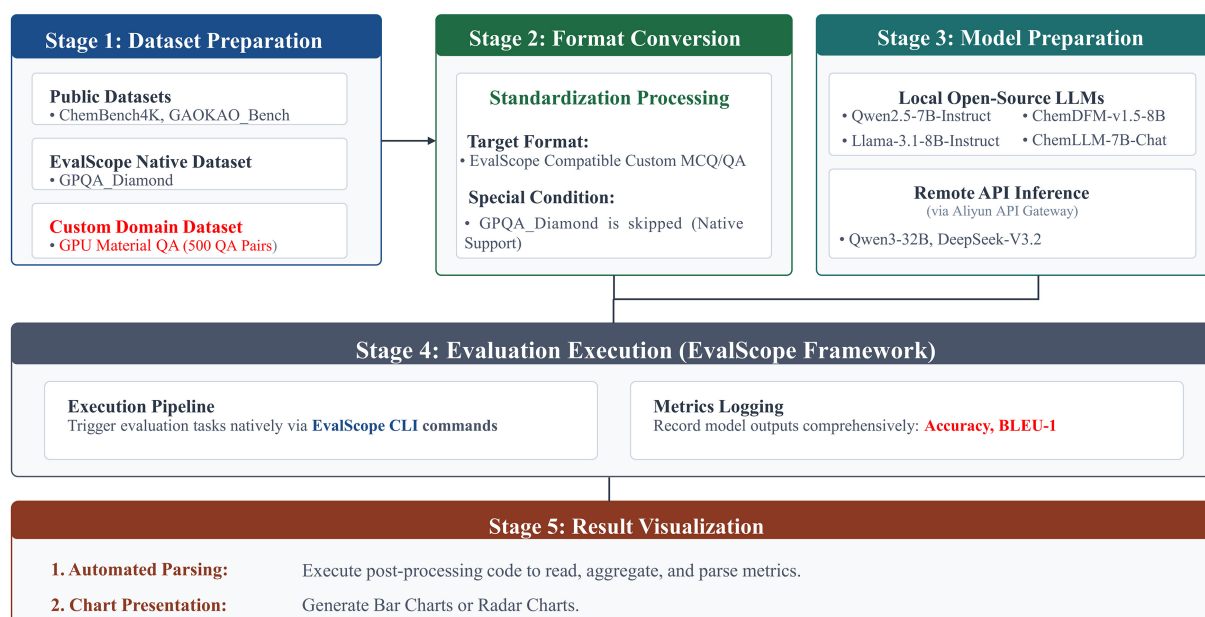


Figure 2. Schematic diagram of the benchmark evaluation process for domain-specific LLMs. Stage 1 (Dataset Preparation): public datasets (ChemBench4K, GAOKAO_Bench), the EvalScope-native GPQA-Diamond, and the custom GPU materials set (500 QA pairs) are assembled. Stage 2 (Format Conversion): the datasets are standardized into the EvalScope-compatible MCQ/QA format, while GPQA-Diamond is skipped as it is natively supported. Stage 3 (Model Preparation): local open-source LLMs (Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, ChemDFM-v1.5-8B, ChemLLM-7B-Chat) and remote API models (Qwen3-32B, DeepSeek-V3.2) are configured. Stage 4 (Evaluation Execution): evaluation is run through the EvalScope framework, recording accuracy for the four MCQ benchmarks and BLEU-1 for the open-ended gpumaterials benchmark. Stage 5 (Result Visualization): the logged metrics are parsed and presented as bar charts. GPU: Graphics processing unit; QA: question-answer; MCQ: multiple-choice question; GPQA: graduate-level Google-proof Q&A; LLMs: large language models; API: application programming interface; CLI: command-line interface; BLEU-1: bilingual evaluation understudy, 1-gram.

RESULTS AND DISCUSSION

Corpus dataset construction

The training corpus of 44,629 samples was assembled from three complementary domain source types together with non-domain supplementary data. The domain topic distribution of the GPU-material QA and MCQ corpus (36,629 domain samples) is shown in the horizontal bar chart in [Figure 3A](#), divided into 21 sub-domain categories based on rule-based keyword matching and sorted by sample size. The categories directly related to GPU thermal management account for over 67% of the samples: semiconductor substrates (18.2%), thermal-management materials (17.4%), packaging, interconnect and interface layers (12.4%), metallic materials (7.5%), TIMs (5.8%), carbon-based materials (2.3%), testing and characterization (1.7%), manufacturing processes (1.2%), and reliability and failure analysis (1.0%). The “cross-domain and methodological” category (19.6%) mainly includes machine-learning and data-driven methods extracted from the methodology sections of GPU-related papers (such as variational autoencoders and materials-informatics pipelines) and semiconductor device-physics content (such as chip-defect detection and nanowire fabrication). Although not directly related to any specific TIM formulation, these topics represent the fundamental computational methods and knowledge of modern GPU-material research, a deliberate design choice that enriches the ability to reason about materials-discovery workflows and manufacturing environments.

The ring chart [[Figure 3B](#)] shows the composition of the entire training corpus. It distinguishes domain-specific QA and MCQ data (82.1%, including QA pairs from GPU-material papers, books, and patents together with synthetic GPU-material MCQs) from non-domain supplementary data (17.9%). The

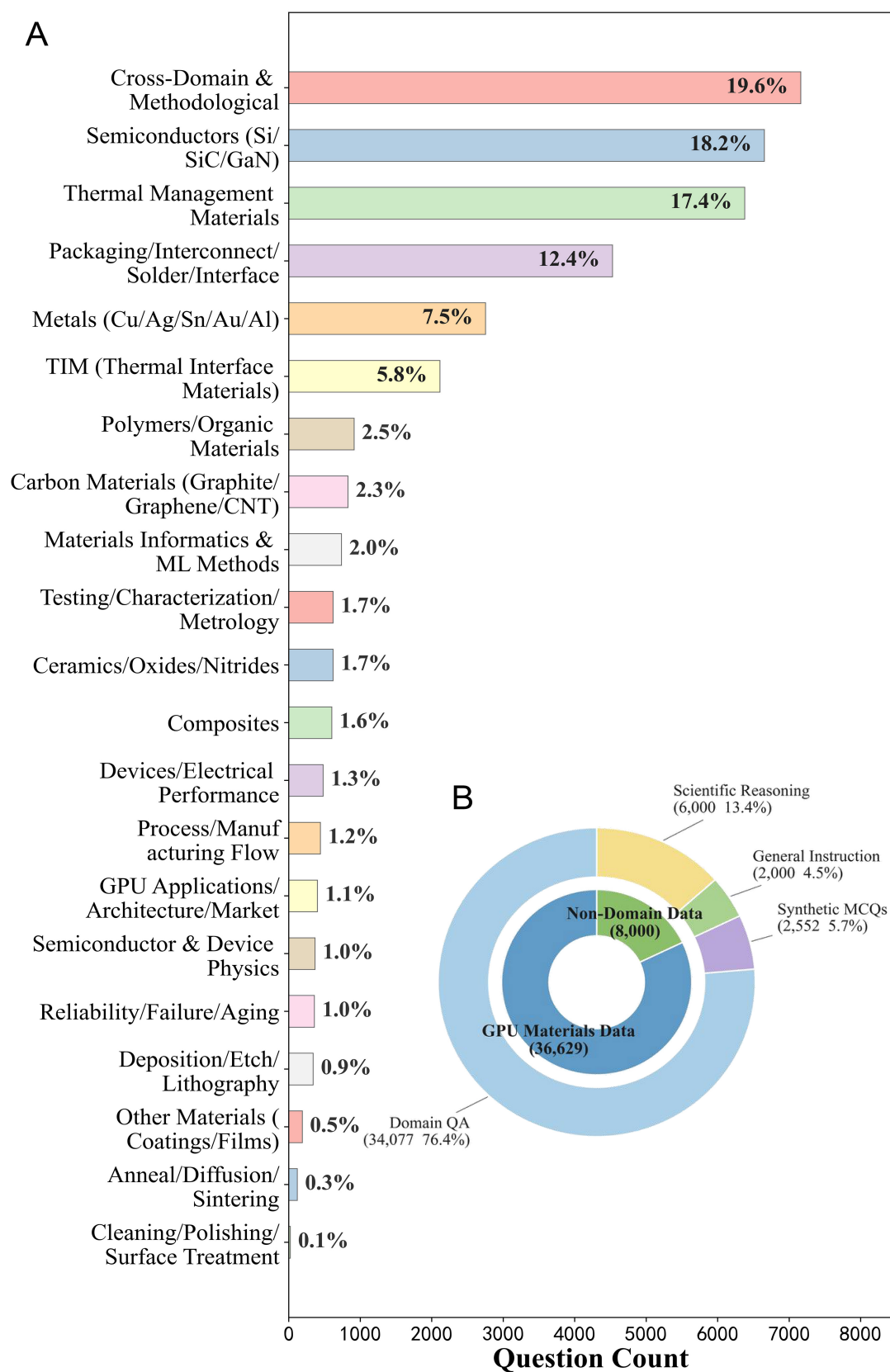


Figure 3. (A) Composition of the GPU materials science training corpus. The horizontal bar chart shows the domain topic distribution of the QA and MCQ data; (B) The layered ring chart summarizes the full 44,629-sample corpus by data type. CNT: carbon nanotube; ML: machine learning; GPU: graphics processing unit; TIM: thermal interface material; QA: question-answer; MCQs: multiple-choice questions.

supplementary data is divided into functional categories to support targeted capability retention and task generalization. General-instruction resources, including 1,000 samples from the Magpie-Qwen2.5-Pro-1M-v0.1 dataset and 1,000 examples from the Firefly Chinese instruction corpus, are leveraged to sustain the fundamental general language competence. Scientific-reasoning samples, consisting of 3,000 entries from the Camel-AI/physics dataset and 3,000 entries from the Microsoft Orca-Math-Word-Problems-200k dataset, maintain the model's capabilities in domain-adjacent logical inference and mathematical reasoning. In addition, 2,552 synthetic GPU-material MCQ samples are incorporated to improve the model's generalization on MCQ-format tasks. The 17.9% non-domain proportion represents a calibrated, intentional mixing ratio that maintains general ability without diluting the domain training signal, and it is reinforced by two independent ability-maintenance mechanisms: data-level mixing and parameter-level freezing (LoRA, with frozen pre-trained weights W_0). The final corpus composition reflects a deliberate blending strategy: 82.1% domain-specific QA and MCQ data provides the targeted knowledge signal, while 17.9% general-instruction and scientific-reasoning data is intended to mitigate catastrophic forgetting and MCQ-format generalization failure.

The gpumaterials evaluation set comprises 500 open-ended QA pairs covering the principal areas of GPU packaging materials science. When grouped into the six sub-domains used for the performance analysis below, the evaluation set shows that packaging and integration accounts for the largest share ($n = 232$) followed by thermal management ($n = 162$), simulation and characterization ($n = 35$), materials and composites ($n = 33$), interconnects and signal integrity ($n = 19$), and processing and manufacturing ($n = 19$). The predominance of packaging and thermal-management items reflects the central role of heat dissipation in contemporary high-power GPU devices. The comparatively small number of items in the remaining four sub-domains is a limitation of the present benchmark: performance estimates for these areas carry correspondingly wider uncertainty, as reflected in the standard deviations reported in the sub-domain analysis below.

Overall, the domain distribution of this dataset reflects the research pattern of GPU materials and packaging technology across subfields. The distribution is broadly consistent with current research activity in industry and academia, and the gpumaterials dataset provides a domain-matched evaluation benchmark for GPU materials and process research.

Benchmark evaluation

Benchmarks on science, chemistry, and physics (MCQ tasks)

This section analyzes the performance of GPUmatLLM and baseline models on MCQ tasks covering multidisciplinary scientific and disciplinary knowledge. The overall benchmark comparison is shown in [Figure 4](#).

Evaluated on the graduate-level multidisciplinary scientific reasoning benchmark GPQA-Diamond, model performance presents distinct hierarchical differences. As illustrated in [Figure 4A](#), the large general-purpose model DeepSeek-V3.2 achieves the best result with a score of 0.7576, followed by Qwen3-32B at 0.5253. The proposed GPUmatLLM reaches a score of 0.3384, which is slightly lower than that of its base model, Qwen2.5-7B-Instruct (0.3636) with a relative degradation of 6.9% (an absolute decrease of 0.0252). This marginal decline suggests a mild attenuation of general scientific reasoning, a trade-off commonly reported in domain-specific fine-tuning. Given the size of the evaluation set, the difference should be regarded as indicative rather than statistically established.

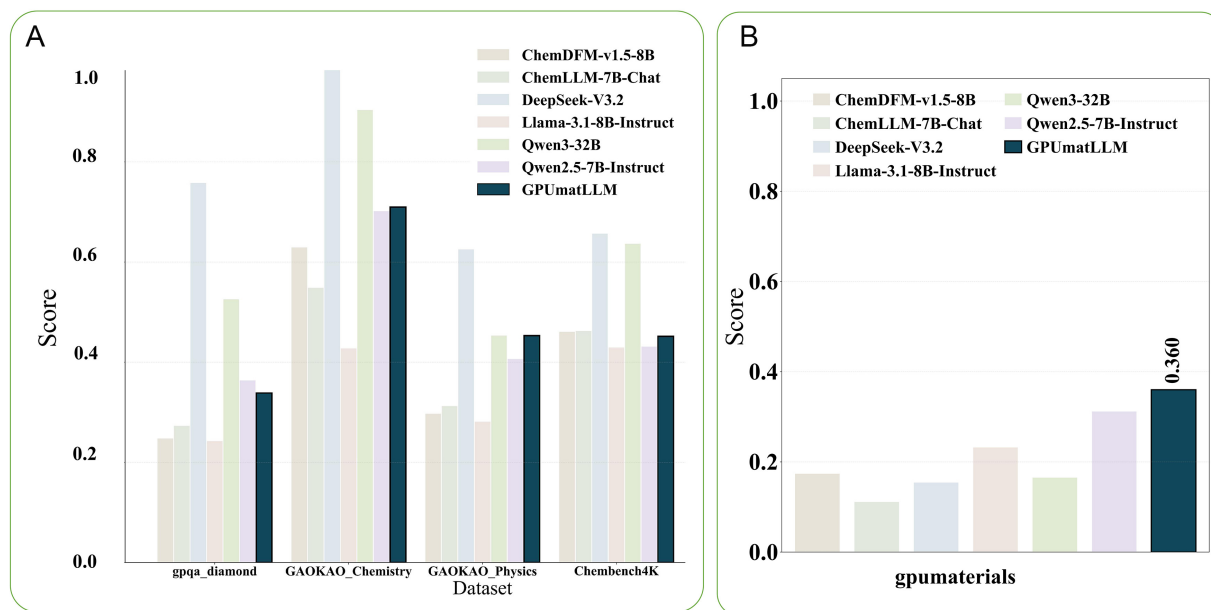


Figure 4. (A) Accuracy of all seven models on the four MCQ benchmarks (GPQA-Diamond, GAOKAO-Chemistry, GAOKAO-Physics, ChemBench4K); (B) BLEU-1 of all seven models on the open-ended gpumaterials benchmark. Accuracy and BLEU-1 are shown in separate panels because the two metrics are not directly comparable. Both panels were generated by the authors using Python with Matplotlib. MCQ: Multiple-choice question.

Chemistry knowledge is further assessed on the GAOKAO-Chemistry benchmark, which covers high-school-level chemistry. DeepSeek-V3.2 and Qwen3-32B achieve the highest scores of 0.9839 and 0.9032, respectively. GPUmatLLM obtains a competitive score of 0.7097, securing the third position among all comparison models. In particular, GPUmatLLM outperforms its base model Qwen2.5-7B-Instruct (0.7016) by an absolute value of 0.0081, and also exceeds professional chemistry models including ChemDFM-v1.5-8B (0.6290) and ChemLLM-7B-Chat (0.5484). These results suggest that domain fine-tuning yields positive transfer on fundamental chemical disciplines without severe knowledge degradation.

Experiments on the GAOKAO-Physics dataset are conducted to examine the physical intuition and problem-solving capacity of the model. DeepSeek-V3.2 again ranks first, with a score of 0.6250, though the performance gap between models is narrowed, revealing a common performance bottleneck in advanced physical reasoning tasks. Notably, GPUmatLLM achieves a score of 0.4531, which is on par with that of the larger-scale Qwen3-32B model and reflects an 11.5% relative improvement compared with its base model Qwen2.5-7B-Instruct (0.4062). This improvement is consistent with cross-disciplinary positive transfer. Given the inherent correlation between materials science and the physical properties of matter (e.g., electrical, optical, and thermal characteristics), the domain-specific corpora introduced during fine-tuning effectively strengthen the core physical reasoning capability of the model.

The comprehensive chemistry benchmark ChemBench4K is adopted to evaluate the breadth and depth of the model's chemical knowledge reserve. DeepSeek-V3.2 (0.6563) and Qwen3-32B (0.6363) remain at the top level of overall performance. GPUmatLLM achieves a score of 0.4515, delivering a 4.8% performance gain over its base model, Qwen2.5-7B-Instruct (0.4310), while slightly underperforming two chemistry-specialized models, ChemLLM-7B-Chat (0.4622) and ChemDFM-v1.5-8B (0.4607). This performance gap stems from the inherent trade-off in domain fine-tuning: unlike dedicated chemistry models optimized for general chemical tasks, GPUmatLLM is fine-tuned to focus on material synthesis and processing scenarios, which naturally limits its generalization ability on generic chemical benchmark datasets.

Evaluation on target domain

The target gpumaterials dataset is designed to evaluate open-ended question-answering capabilities focusing on materials science and processing scenarios. Different from the aforementioned MCQ tasks, this benchmark places greater demands on domain knowledge retrieval and on accurate, coherent text generation, posing a substantial challenge for existing models.

The quantitative results visualized in [Figure 4B](#) show that GPUmatLLM achieves the highest BLEU-1 score among the evaluated models (0.3599), ahead of all baselines on this domain-specific open-ended task. Further quantitative analysis yields several key findings from the benchmark comparisons.

The proposed model delivers significant performance improvement over its base architecture. GPUmatLLM outperforms Qwen2.5-7B-Instruct (0.3115) by an absolute margin of 0.0484, corresponding to a 15.5% relative performance gain. This represents the most notable improvement across all tested benchmarks, which supports the effectiveness of LoRA-based fine-tuning on the customized GPU materials processing dataset constructed in this work.

GPUmatLLM also shows a clear advantage over mainstream general-purpose large models. General models including DeepSeek-V3.2 (0.1541) and Qwen3-32B (0.1650) achieve superior results on conventional MCQ benchmarks yet experience substantial performance degradation on this open-ended materials task, with the score of GPUmatLLM more than doubling that of DeepSeek-V3.2. Such a marked performance discrepancy reflects the steep domain-knowledge barrier inherent to materials science. Increasing model scale alone, without targeted domain alignment, does not confer competitive performance in vertically specialized application scenarios.

GPUmatLLM also exhibits superiority over peer models of comparable scale and domain-specialized chemistry models. The proposed model outperforms the general-scale Llama-3.1-8B-Instruct (0.2319), as well as chemistry-optimized models including ChemDFM-v1.5-8B (0.1736) and ChemLLM-7B-Chat (0.1110). This outcome indicates that generic chemistry domain adaptation cannot be directly generalized to support high-performance reasoning in this specialized subfield of materials science.

All evaluated models yield relatively low overall scores below 0.40 on the gpumaterials benchmark, revealing the inherent difficulty of open-ended QA tasks in materials science. While GPUmatLLM achieves the best performance among all comparison models in this task, the limited overall scores suggest considerable room for subsequent optimization. Overall, GPUmatLLM realizes efficient domain specialization, delivering superior domain-specific performance with a lightweight 7B-parameter architecture, outperforming large general-purpose LLMs.

When evaluated on specific benchmarks in the GPU materials science field, GPUmatLLM produced higher scores compared to its base model (Qwen2.5-7B-Instruct), as shown in [Figure 5A](#) and [B](#). On the domain-specific gpumaterials benchmark, GPUmatLLM outperforms its base model by a relative 15.5%, confirming the effect of domain-specific fine-tuning. Both models perform very similarly on three general scientific benchmarks, including ChemBench4K, GAOKAO-Physics, and GAOKAO-Chemistry, indicating that fine-tuning has only a marginal effect on the general scientific reasoning ability of the model.

GPUmatLLM and Qwen2.5-7B-Instruct show comparably low accuracy on the GPQA-Diamond benchmark, reaching 0.3384 and 0.3636, respectively. The marginal performance difference indicates that fine-tuning on GPU-materials data has almost no impact on model performance in this out-of-domain scenario. On the GAOKAO-Chemistry dataset, both GPUmatLLM and Qwen2.5-7B-Instruct achieved their highest scores

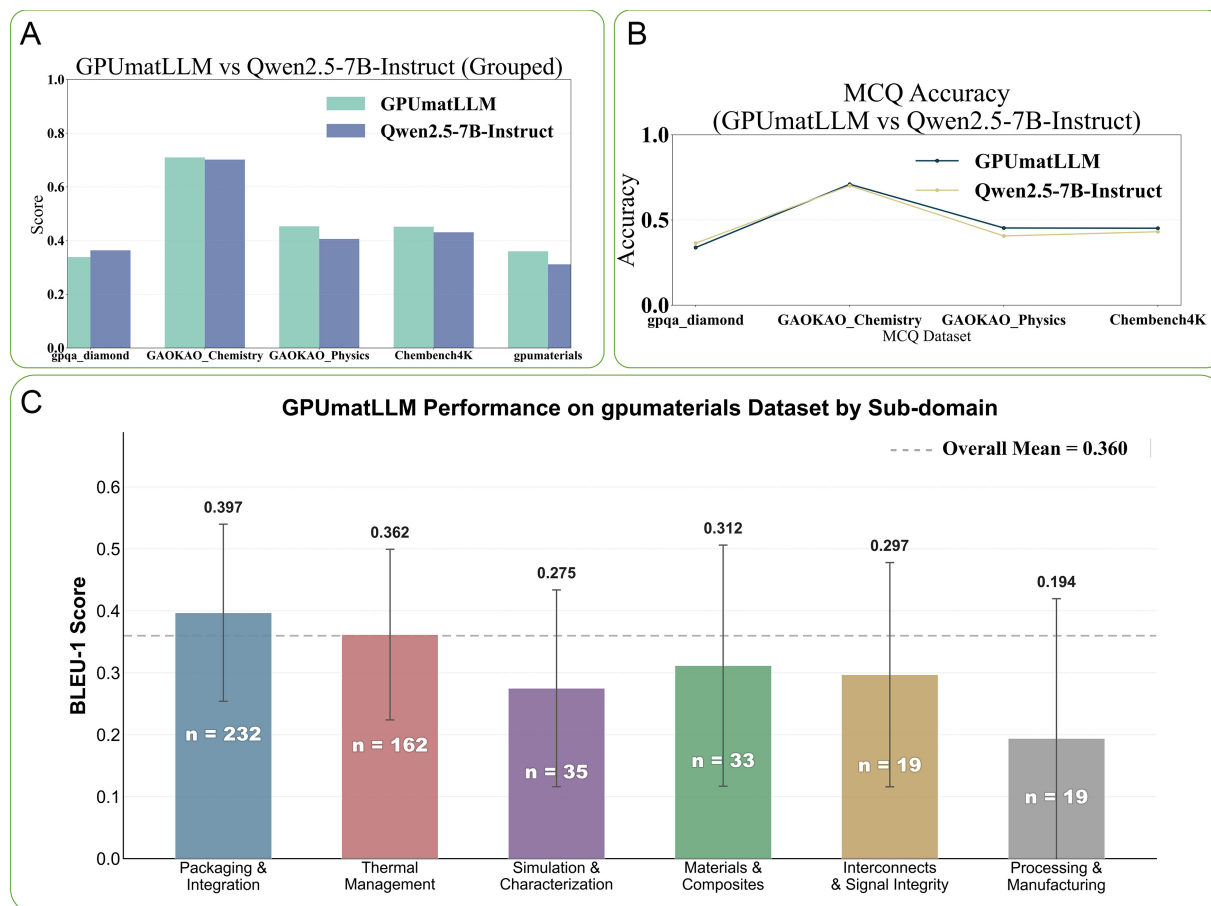


Figure 5. (A) Grouped bar chart comparing GPUmatLLM with its base model Qwen2.5-7B-Instruct across the five evaluation sets; accuracy is reported for the four MCQ benchmarks and BLEU-1 for gpumaterials; (B) Line chart comparing MCQ accuracy between GPUmatLLM (blue line) and Qwen2.5-7B-Instruct (yellow line) across four MCQ benchmarks: GPQA-Diamond, GAOKAO-Chemistry, GAOKAO-Physics, and ChemBench4K (x-axis, left to right). The y-axis denotes classification accuracy on a 0-1 scale; (C) GPUmatLLM performance on the gpumaterials benchmark by sub-domain; error bars denote one standard deviation, and n is the number of items per sub-domain. All panels were generated by the authors using Python with Matplotlib. MCQ: Multiple-choice question; BLEU-1: bilingual evaluation understudy, 1-gram; GPQA: graduate-level Google-proof Q&A.

among the four MCQ benchmarks (0.7097 and 0.7016, respectively), with GPUmatLLM exhibiting a slight advantage that suggests a positive transfer effect from chemical domain knowledge acquired during fine-tuning. On the GAOKAO-Physics benchmark, the accuracy of GPUmatLLM was 0.4531, compared with 0.4062 for Qwen2.5-7B-Instruct. The proposed model nevertheless maintains a slight performance advantage. The test results on ChemBench4K show that the scores between the two models are very similar, with GPUmatLLM of 0.4515 and Qwen2.5-7B-Instruct of 0.4310. This confirms that fine-tuning does not cause measurable performance regression in general chemical multiple-choice tasks. The closely matched performance across the four MCQ benchmarks indicates that LoRA fine-tuning on the GPU-materials corpus largely preserves the general scientific reasoning ability of the base model, with a slight regression on GPQA-Diamond (-0.0252) and gains on the remaining three benchmarks. We note that the individual contributions of the 17.9% auxiliary-data mixing and of LoRA parameter freezing are not separated in the present work; a dedicated ablation isolating the two mechanisms is left for future study.

Figure 5C shows the BLEU-1 score distribution of GPUmatLLM across the six sub-domains of the gpumaterials dataset. The model has an overall mean BLEU-1 score of 0.3599, and its performance differs markedly across sub-domains. The packaging and integration sub-domain receives the highest score (0.3970) and has the largest sample size ($n = 232$), suggesting the model's generative ability is relatively stable for this

type of problem. The score for the thermal management sub-domain is 0.362, higher than the overall mean, reflecting the model's grasp of knowledge related to heat dissipation and conduction. The materials and composites sub-domain scores 0.312, but with a standard deviation of 0.195, showing that performance varies widely in this sub-domain; answers to some materials science questions are of relatively high quality, whereas those involving specific characterization data or formula details are inaccurate. Scores for the interconnects and signal integrity (0.2970) and simulation and characterization (0.2750) sub-domains are both below the overall mean, pointing to knowledge gaps in areas such as electrical interconnection design, high-precision measurement, and modeling. The processing and manufacturing sub-domain has the lowest score (0.1940) and the largest standard deviation (0.2260), suggesting that the model lacks accuracy in answering questions about specific process parameters such as deposition, etching, and solidification, a limitation that should be addressed in future work.

In terms of data distribution, the two sub-domains of packaging and integration, and thermal management account for approximately 79% of all samples and thus constitute the bulk of the dataset. The model's performance in these two dominant sub-domains strongly supports the overall score; however, sub-domains with smaller sample sizes generally show larger standard deviations, indicating that the reliability of the model's performance under small-sample conditions needs further validation.

We extracted four representative GPUmatLLM responses to domain-specific questions from the GPU material evaluation set [Figure 6], demonstrating the model's ability to produce mechanistically grounded answers using precise domain terminology in the core subdomains of the training corpus. The problem shown in Figure 6A mainly concerns the miniaturization of semiconductor packaging. GPUmatLLM correctly identifies the roles of spatial integration based on interposers, placement of TIMs between chips and substrates, metallization layers for electrical interconnects, and three-dimensional stacking in achieving device miniaturization, demonstrating multi-level packaging knowledge across material selection, thermal management, and mechanical architecture. Figure 6B is about the mixed bonding problem in chip stacking. GPUmatLLM accurately describes mixed bonding as the simultaneous direct bonding of metal to metal and dielectric to dielectric, elucidating the roles of metallization layers and insulation layers in forming metallurgical and chemical/physical bonds, and correctly linking this technology with shorter inter-chip distances, reduced signal delay, lower power consumption, and higher packaging density in vertical chip integration. Figure 6C shows the amorphous carbon interlayer thermal modulation problem. GPUmatLLM proposes a mechanism-level explanation that links the insertion of amorphous carbon interlayers at the copper-diamond interface with the suppression of phonon scattering, promotion of phonon transport, enhancement of interface bonding strength, and overall improvement of interface thermal conductivity. Answering this question requires comprehensive knowledge of phonon transport physics and interface materials science. Figure 6D shows the issue of protrusions and indentations in the heat sink. GPUmatLLM correctly attributes the functional role of surface features on the heat sink base to airflow guidance, convective heat transfer enhancement, compression of TIMs, and stress release at the printed circuit board (PCB) interface, demonstrating an application-level understanding of GPU thermal management hardware. In all four cases, GPUmatLLM demonstrated accurate use of domain-specific terminology (highlighted in orange and blue in the original annotations), process-oriented causal reasoning, and knowledge integration across sub-domains of materials, heat, and packaging engineering.

Representative examples from the evaluation set are provided to further illustrate the qualitative advantages of GPUmatLLM. In Figure 7A, GPUmatLLM provides a structured and technically detailed explanation of via interconnects, covering vertical electrical connectivity, 3D network construction, and signal integrity, whereas Llama-3.1-8B-Instruct produces only a generic two-sentence response lacking domain-specific terminology. In example Figure 7B, while both models correctly identify copper and aluminum as typical

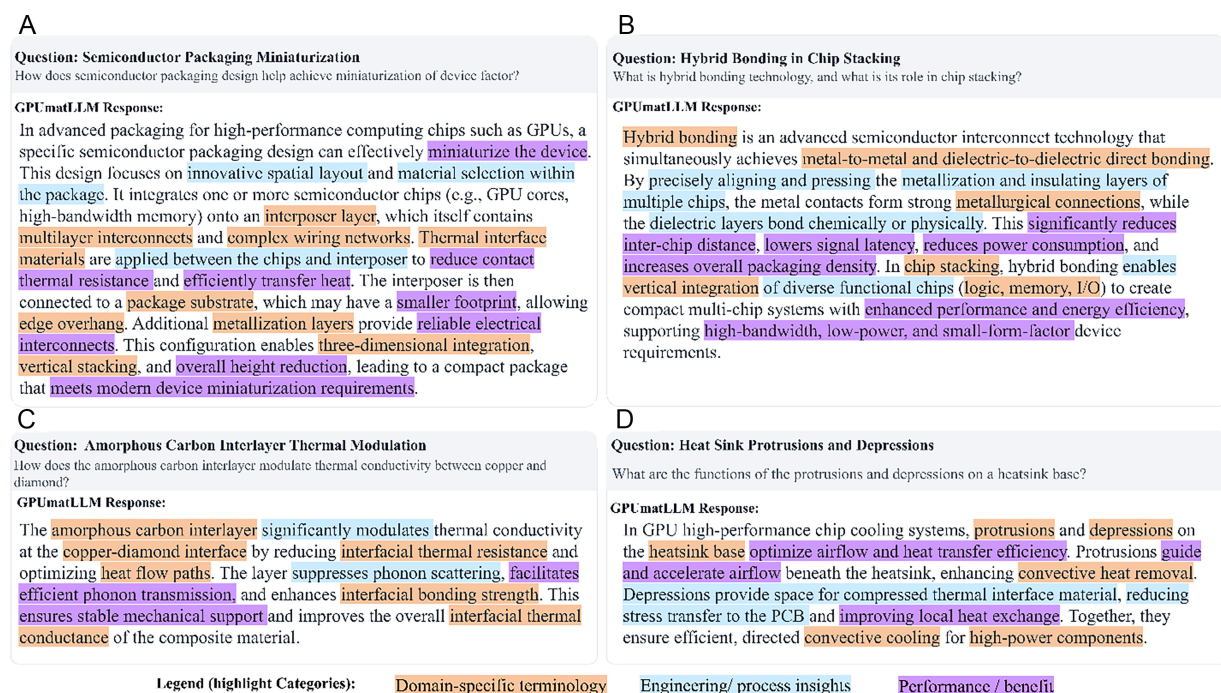


Figure 6. Four representative GPUmatLLM responses to domain-specific questions drawn from the GPU materials science evaluation set (Orange represents Domain-specific terminology, blue represents Engineering/Process Insights, and purple represents Performance/Benefit). GPU: Graphics processing unit; PCB: printed circuit board; I/O: input/output.

thermal module materials, GPUmatLLM additionally describes the heat transfer pathway from chip to heat spreader fins, demonstrating a deeper understanding of thermal management mechanisms. These examples suggest that GPUmatLLM, fine-tuned on GPU material and packaging domain data, generates responses with greater technical depth and structural completeness than general-purpose models of similar scale. **Figure 7C** and **D** present a second qualitative comparison against ChemLLM-7B-Chat, a domain-specialized chemistry model of similar scale. In **Figure 7C**, GPUmatLLM provides a structured, multi-aspect account of heat-spreader functions, identifying the low-thermal-resistance conduction path, the lateral heat spreading that lowers the chip junction temperature, the mechanical support and stress buffering against coefficient-of-thermal-expansion (CTE) mismatch, and the compatibility with external cooling solutions such as heat sinks, vapor chambers, and cold plates. In contrast, ChemLLM-7B-Chat returns a single generic sentence stating that heat spreaders serve a heat-dissipation purpose, without a packaging-specific mechanism. In **Figure 7D**, GPUmatLLM explains that metallic supporting structures establish a low-thermal-resistance pathway from the heat source to the external cooling system and that their high conductivity reduces the heat-transmission resistance along this path, explicitly linking the design to stable GPU operation under high load; ChemLLM-7B-Chat again offers only a generic statement about transferring heat to a heat sink, omitting the GPU packaging context, the thermal-interface coupling, and the low-thermal-resistance rationale. There are two more examples attached in **Supplementary Figures 3** and **4**. These examples indicate that domain-specific fine-tuning yields responses with substantial technical depth and structural completeness.

Limitations

Two limitations should be kept in mind when interpreting the reported scores. First, the benchmark is model-generated and automatically validated rather than expert-authored, so the scores reflect agreement with an automatically derived reference, and the benchmark inherits any systematic bias from the generating model. Second, the benchmark is drawn from the same literature domain as the training corpus, so it probes in-domain knowledge retention rather than generalization to unseen sub-domains. Expert-authored



Figure 7. (A and B) Qualitative comparison of responses generated by GPUmatLLM and Llama-3.1-8B-Instruct on GPU packaging domain questions with (A) via interconnect function in advanced packaging, and (B) thermal module materials and heat dissipation mechanisms; (C and D) Qualitative comparison of responses generated by GPUmatLLM and ChemLLM-7B-Chat on GPU packaging and thermal-management questions; (C) Main functions of heat spreaders in semiconductor packaging; (D) Purpose of adopting thermally conductive metals for the supporting structure. GPU: Graphics processing unit; RDL: redistribution layer; TIM: thermal interface material; CTE: coefficient-of-thermal-expansion; PCB: printed circuit board; I/O: input/output.

reference answers, semantic scoring metrics such as BERTScore, and an out-of-distribution evaluation split are natural next steps and are left for future work.

CONCLUSIONS

This study establishes a dedicated benchmark for GPU materials science, comprising a systematically constructed evaluation dataset and a comprehensive assessment framework for LLMs. To support this architecture, we compiled a training corpus of 44,629 samples, of which 36,629 are domain data derived from academic papers, technical books, and industrial patents (34,077 QA pairs and 2,552 synthesized MCQs), complemented by 8,000 general-domain auxiliary samples. Although frontier general-purpose models such as DeepSeek-V3.2 and Qwen3-32B lead on general scientific benchmarks, they fail to capture GPU material-specific knowledge. Similarly, chemistry-focused models such as ChemLLM-7B-Chat and ChemDFM-v1.5-8B perform strongly on general chemistry tests but show inadequate accuracy on our specialized GPU materials tasks.

Building upon this benchmark, we developed GPUmatLLM, a specialized model fine-tuned on the same corpus, to validate the benchmark's effectiveness and explore domain-specific adaptation. By deploying parameter-efficient LoRA fine-tuning on the Qwen2.5-7B-Instruct backbone, our model achieved a notable 15.5% performance boost over its base version, and consistently surpassed six mainstream baseline models in domain knowledge while retaining strong general scientific capabilities. These results indicate that in specialized engineering domains, targeted domain knowledge can matter more than model scale alone.

Future work will expand our training corpus with practical industrial datasets and technical datasheets to broaden knowledge coverage. We also plan to conduct comprehensive comparative evaluations against more recent materials models to further position our work in the field. Finally, we will explore continuous domain pre-training alongside LoRA optimization to enhance model reliability, reduce errors, and improve its readiness for real-world industrial deployment. In summary, this work represents a systematic attempt to develop a dedicated benchmark and LLM for GPU-related materials science.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study, performed corpus construction, model fine-tuning, evaluation and data analysis, and drafted the manuscript: Wang, Y. (Yuqin Wang)
Performed data curation and validation and revised the manuscript critically for important intellectual content: Wang, Y. (Youle Wang)
Made substantial contributions to the conception of the study, provided supervision, resources and funding acquisition, and revised the manuscript: Zhang, L.
All authors read and approved the final version of the manuscript.

Availability of data and materials

The fine-tuned GPUmatLLM model, the training corpus, and the gpumaterials evaluation set are publicly available. The model is released at <https://huggingface.co/wvss/GPUmatLLM>; the training dataset and the gpumaterials benchmark are available at <https://huggingface.co/datasets/wvss/GPUmatLLM-train> and <https://huggingface.co/datasets/wvss/GPUmatLLM-eval>, respectively.

AI and AI-assisted tools statement

Large language models were used as objects of study and as data-generation tools in this work. Their use is integral to the research and is described in full in the Methods section, covering corpus construction, question-answer and multiple-choice generation, automated quality scoring, and benchmark evaluation.

During the preparation of this manuscript, the AI tool DeepSeek-V4-Flash (version 0731, released 2026-07-31) was used solely for language editing. The tool did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This work was supported by the Chinese Academy of Sciences Pioneer Initiative (CAS Leading Action Plan) (Grant No. E65543Z1).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

Supplementary Materials

[Supplementary Materials](#)

REFERENCES

- Ahlawat, D.; Mishra, V.; Singh, S.; et al. A family of large language models for materials research with insights into model adaptability in continued pretraining. *Nat. Mach. Intell.* **2026**, *8*, 435–48. [DOI](#)
- Mandal, I.; Soni, J.; Zaki, M.; et al. Evaluating large language model agents for automation of atomic force microscopy. *Nat. Commun.* **2025**, *16*, 9104. [DOI PubMed PMC](#)
- Zhu, Y.; Yuan, H.; Wang, S.; et al. Large language models for information retrieval: a survey. *ACM. Trans. Inf. Syst.* **2025**, *44*, 1–54. [DOI](#)
- Lv, S.; Peng, L.; Wu, W.; Yao, Y.; Jiao, S.; Hu, W. Bridging language models and computational materials science: A prompt-driven framework for material property prediction. *Mater. Genome. Eng. Adv.* **2025**, *3*, e70013. [DOI](#)
- Truhn, D.; Reis-Filho, J. S.; Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat. Med.* **2023**, *29*, 2983–4. [DOI PubMed](#)
- Asai, A.; He, J.; Shao, R.; et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature* **2026**, *650*, 857–63. [DOI PubMed PMC](#)
- Xiao, J.; Sun, Z.; An, H.; Zhao, H.; Qiu, M.; Li, X. Optical image processing and applications empowered by vision-language models. *iOptics* **2025**, *1*, 100003. [DOI](#)
- Zeng, Y.; Jin, X. AI-empowered optical metasurfaces: From unit-cell optimization to system-level integration. *iOptics* **2025**, *1*, 100004. [DOI](#)
- Zang, Y.; Zhang, Z.; Li, S.; Zhang, F.; Chen, H. Fiber neural networks for the intelligent optical fiber communication signal processing. *iOptics* **2025**, *1*, 100009. [DOI](#)
- Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *arXiv* **2023**, arXiv:1706.03762. Available online: <https://doi.org/10.48550/arXiv.1706.03762> (accessed 15 October 2026).
- DeepSeek-V3 technical report. *arXiv* **2025**, arXiv:2412.19437. Available online: <https://doi.org/10.48550/arXiv.2412.19437> (accessed 15 October 2026).
- He, J.; Guan, H.; Feng, W.; et al. Controlling risks of AI in chemical science with agents. *AI. Sci.* **2025**, *1*, 015002. [DOI](#)
- Wei, B.; Luo, W.; Du, J.; et al. Thermal interface materials: From fundamental research to applications. *SusMat* **2024**, *4*, e239. [DOI](#)
- Katari, M.; Jeyaraman, J.; Mohamed, I. A.; Kumaran, T. Addressing power and thermal challenges in advanced packaging for AI CPUs/GPUs: insights into multi-die stacking technology. *Int. J. Multidiscip. Res.* **2023**, *5*, 17017. [DOI](#)
- Xu, K.; Lu, Y.; Liu, L.; Zhang, L. MatterMind: a user-friendly material design platform assisted by large language model. *Mater. Genome. Eng. Adv.* **2026**, *4*, e70075. [DOI](#)
- Adachi, Y.; Chen, T. T.; Sun, F.; et al. A review on inverse analysis models in steel material design. *Mater. Genome. Eng. Adv.* **2024**, *2*, e71. [DOI](#)
- Wang, Z.; Wu, Z.; Weng, L.; et al. A roadmap review of thermally conductive polymer composites: critical factors, progress, and prospects. *Adv. Funct. Mater.* **2023**, *33*, 2301549. [DOI](#)
- Tolstaya, A.; Suslina, I.; Tolstaya, P. Review of information databases providing data on current scientific and technical achievements. *Procedia. Comput. Sci.* **2016**, *88*, 385–90. [DOI](#)
- Austin, T. Towards a digital infrastructure for engineering materials data. *Mater. Discov.* **2016**, *3*, 1–12. [DOI](#)
- Zhang, L.; Deng, H. NJmat 2.0: user instructions of data-driven machine learning interface for materials science. *Comput. Mater. Contin.* **2025**, *83*, 1–11. [DOI](#)
- Zhang, L.; He, M.; Shao, S. Machine learning for halide perovskite materials. *Nano. Energy.* **2020**, *78*, 105380. [DOI](#)
- Agrawal, A.; Choudhary, A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. *APL. Mater.* **2016**, *4*, 053208. [DOI](#)
- Cole, J. M. A design-to-device pipeline for data-driven materials discovery. *Acc. Chem. Res.* **2020**, *53*, 599–610. [DOI PubMed](#)
- Li, S.; Zhou, X.; Liu, Y.; et al. Agile conceptual design and validation based on multi-source product data and large language models: a review, framework, and outlook. *J. Eng. Des.* **2025**, *36*, 473–503. [DOI](#)
- Lin, Z.; Yang, C. Artificial intelligence design of sustainable aluminum alloys: a review. *Comput. Mater. Contin.* **2026**, *86*, 1–33. [DOI](#)
- Zhang, L.; Wang, Y.; Zhou, Y.; Li, C. AI-guided design of optoelectronic molecularly modified CH₃NH₃PbI₃ perovskite thin films with improved aqueous stability. *AI. Sci.* **2026**, *2*, 025006. [DOI](#)
- Zhang, Q.; Zhuang, X.; Zhou, C.; et al. Beyond accuracy: comprehensive alignment for AI-driven molecular design. *AI. Sci.* **2026**, *2*, 023001. [DOI](#)
- Yao, T.; Xu, A.; Ou, P.; Yang, W. Let materials science data learn to reason. *AI. for. Mater.* **2025**, *1*, 4. [DOI](#)
- Cao, B.; Yang, S.; Sun, A.; Dong, Z.; Zhang, T. Y. Domain knowledge-guided interpretive machine learning: formula discovery for the oxidation behavior of ferritic-martensitic steels in supercritical water. *J. Mater. Inf.* **2022**, *2*, 4. [DOI](#)

30. Lu, T.; Li, M.; Lu, W.; Zhang, T. Y. Recent progress in the data-driven discovery of novel photovoltaic materials. *J. Mater. Inf.* **2022**, *2*, 7. DOI
31. Li, L.; Xu, K.; Su, R.; Gu, H.; Ma, P. Data speaks: LLM-enabled evidence auditing resets the “migration barrier” playbook for solid-state electrolytes. *AI. for. Mater.* **2025**, *1*, 2. DOI
32. Tshitoyan, V.; Dagdelen, J.; Weston, L.; et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95-8. DOI PubMed
33. Lee, J.; Jung, W.; Baek, S. In-house knowledge management using a large language model: focusing on technical specification documents review. *Appl. Sci.* **2024**, *14*, 2096. DOI
34. Ralph, J.; Martynov, P.; Ma, X.; et al. Identifier service in the Mindat database: persistent and structured access to massive records of minerals and other natural materials. *Data. Intell.* **2025**, *7*, 692-711. DOI
35. Luo, Y.; Liu, Y.; Zhang, L.; Gao, F.; Gu, J. A survey on quality evaluation of instruction fine-tuning datasets for large language models. *Data. Intell.* **2025**, *7*, 527-66. DOI
36. He, F.; Lai, H.; Liu, J.; et al. Solving mathematical problems using large language models: a survey. *Data. Intell.* **2025**, *7*, 907-46. DOI
37. Zhang, L.; Zhou, J.; Chen, X. Data-driven exploration and first-principles analysis of perovskite material. *J. Mater. Inf.* **2024**, *4*, 13. DOI
38. Dagdelen, J.; Dunn, A.; Lee, S.; et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15*, 1418. DOI PubMed PMC
39. Miret, S.; Krishnan, N. M. A. Enabling large language models for real-world materials discovery. *Nat. Mach. Intell.* **2025**, *7*, 991-8. DOI
40. Zhang, S.; Wang, W. Y.; Wang, X.; et al. Large language models enabled intelligent microstructure optimization and defects classification of welded titanium alloys. *J. Mater. Inf.* **2024**, *4*, 34. DOI
41. Umer, M. S.; Nabeel, M.; Athar, U.; et al. Large language models meet molecules: a systematic review of advances and challenges in AI-driven cheminformatics. *Arch. Computat. Methods. Eng.* **2026**, *33*, 4867-908. DOI
42. Zhang, L.; Liu, Z.; Ni, B.; Wang, Q. Large language models (LLMs) for materials design. *Adv. Funct. Mater.* **2025**, *36*, e25897. DOI
43. Zhang, D.; Jia, X.; Wang, Y.; et al. Digital materials ecosystem: from databases to AI agents for autonomous discovery. *Chem. Sci.* **2026**, *17*, 5782-804. DOI PubMed PMC
44. Yan, Z.; Liang, H.; Wang, J.; et al. PDGPT: a large language model for acquiring phase diagram information in magnesium alloys. *Mater. Genome. Eng. Adv.* **2024**, *2*, e77. DOI
45. Hu, Z.; You, J.; Huang, X.; Huang, H.; Tao, J.; Yi, J. Automatic construction of knowledge graphs from scientific literature for copper-based composites. *Data. Intell.* **2024**, *6*, 1168-89. DOI
46. Velivela, P. T.; Zhao, Y. F. BIKAS: bio-inspired knowledge acquisition and simulacrum - a knowledge database to support multifunctional design concept generation. *Data. Intell.* **2024**, *6*, 531-58. DOI
47. Jiang, X.; Wang, W.; Tian, S.; Wang, H.; Lookman, T.; Su, Y. Applications of natural language processing and large language models in materials discovery. *npj. Comput. Mater.* **2025**, *11*, 79. DOI
48. Hu, H.; Stirrat, H. J.; Alayli, A.; Saeki, A.; Huang, Y. AI-powered workflow for constructing organic materials databases from the literature: integrating large language models. *ACS. Omega.* **2025**, *10*, 49545-56. DOI PubMed PMC
49. Liu, C.; Kang, Y.; Wang, S.; et al. More than catastrophic forgetting: integrating general capabilities for domain-specific LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November, 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024, pp 7531-48. DOI
50. Li, X.; Ren, W.; Qin, W.; Wang, L.; Zhao, T.; Hong, R. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, April 6-11, 2025; IEEE, 2025, pp 1-5. DOI
51. Szabó, B. N.; Gyöngyösy, N. M.; Dalos, A. Applied domain adaptation of LLM-based document embeddings for engineering knowledge retrieval. In *IntRob '25: Proceedings of the Intelligent Robotics FAIR 2025*; ACM: New York, NY, USA, 2025, pp 29-37. DOI
52. Guo, T.; Guo, K.; Nan, B.; et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems 36*, New Orleans, Louisiana, USA, December 10-16, 2023; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, California, USA, 2023, pp 59662-88. DOI
53. Liang, P.; Bommasani, R.; Lee, T.; et al. Holistic evaluation of language models. *arXiv* **2022**, arXiv:2211.09110. Available online: <https://doi.org/10.48550/arXiv.2211.09110> (accessed 15 October 2026).
54. Wang, Y.; Wang, Y.; Zhang, L. Benchmarking and fine-tuning large language models for GPU materials science. *ChemRxiv.* **2026**. DOI
55. Tong, X. C. *Advanced Materials for Thermal Management of Electronic Packaging*. Springer Series in Advanced Microelectronics, Vol. 30; Springer New York, 2011. DOI
56. *Thermal Management of Electronic Systems II: Proceedings of EUROTHERM Seminar 45, 20-22 September 1995, Leuven, Belgium*; Beyne, E., Lasance, C. J. M., Berghmans, J., Eds.; Springer Netherlands, 1997. DOI

57. Xu, Z.; Jiang, F.; Niu, L.; et al. Magpie: alignment data synthesis from scratch by prompting aligned LLMs with nothing. *arXiv* **2024**, arXiv:2406.08464. Available online: <https://doi.org/10.48550/arXiv.2406.08464> (accessed 15 October 2026).
58. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: communicative agents for “Mind” exploration of large language model society. In *Advances in Neural Information Processing Systems 36*, New Orleans, Louisiana, USA, December 10–16, 2023; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, California, USA, 2023, pp 51991–2008. DOI
59. Mitra, A.; Khanpour, H.; Rosset, C.; Awadallah, A. Orca-Math: unlocking the potential of SLMs in grade school math. *arXiv* **2024**, arXiv:2402.14830. Available online: <https://doi.org/10.48550/arXiv.2402.14830> (accessed 15 October 2026).
60. Augenstein, I.; Baldwin, T.; Cha, M.; et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **2024**, *6*, 852–63. DOI
61. Miao, Z.; Sun, Q.; Wang, J.; et al. Easy dataset: a unified and extensible framework for synthesizing LLM fine-tuning data from unstructured documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Suzhou, China, November, 2025; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025, pp 960–8. DOI
62. Zheng, Y.; Zhang, R.; Zhang, J.; Yeyanhan, Y.; Luo, Z. LlamaFactory: unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, August, 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024, pp 400–10. DOI
63. Hu, E. J.; Shen, Y.; Wallis, P.; et al. LoRA: Low-Rank Adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685. Available online: <https://doi.org/10.48550/arXiv.2106.09685> (accessed 15 October 2026).
64. Qwen; Yang, A.; Yang, B.; et al. Qwen2.5 technical report. *arXiv* **2025**, arXiv:2412.15115. Available online: <https://doi.org/10.48550/arXiv.2412.15115> (accessed 15 October 2026).
65. Grattafiori, A.; Dubey, A.; Jauhri, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783. Available online: <https://doi.org/10.48550/arXiv.2407.21783> (accessed 15 October 2026).
66. Zhao, Z.; Ma, D.; Chen, L.; et al. Developing ChemDFM as a large language foundation model for chemistry. *Cell. Rep. Phys. Sci.* **2025**, *6*, 102523. DOI
67. Zhang, D.; Liu, W.; Tan, Q.; et al. ChemLLM: a chemical large language model. *arXiv* **2024**, arXiv:2402.06852. Available online: <https://doi.org/10.48550/arXiv.2402.06852> (accessed 15 October 2026).
68. Yang, A.; Li, A.; Yang, B.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388. Available online: <https://doi.org/10.48550/arXiv.2505.09388> (accessed 15 October 2026).
69. DeepSeek-AI; Liu, A.; Mei, A.; et al. DeepSeek-V3.2: pushing the frontier of open large language models. *arXiv* **2025**, arXiv:2512.02556. Available online: <https://doi.org/10.48550/arXiv.2512.02556> (accessed 15 October 2026).
70. Zhang, X.; Li, C.; Zong, Y.; et al. Evaluating the performance of large language models on GAOKAO benchmark. *arXiv* **2023**, arXiv:2305.12474. Available online: <https://doi.org/10.48550/arXiv.2305.12474> (accessed 15 October 2026).
71. Rein, D.; Hou, B. L.; Stickland, A. C.; et al. GPQA: a graduate-level google-proof Q&A benchmark. *arXiv* **2023**, arXiv:2311.12022. Available online: <https://doi.org/10.48550/arXiv.2311.12022> (accessed 15 October 2026).

Disclaimer/Publisher’s Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.