GLASSBENCH: BENCHMARKING LARGE LANGUAGE MODELS FOR DOMAIN-SPECIFIC GLASS MATERIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have shown substantial potential for scientific question answering, reasoning, and knowledge integration, yet materials science still lacks standardized benchmarks for assessing domain-specific capabilities. This limitation is especially pronounced in glass science, where terminology is specialized, composition-structure-property relations are highly nonlinear, and most available resources are structured numerical databases rather than natural-language evaluation sets. Here, we introduce GlassBench, a multitask benchmark for evaluating LLMs in glass materials science. Based on this, GMatLLM is constructed as a domain-specific LLM for the glass research. First, GlassBench was constructed by extracting question-answer pairs from authoritative textbooks and research literature, followed by multistage quality control combining LLM-based scoring, confident-learning-based screening, domain-consistency checks, and expert review. The resulting dataset contains 19,974 quality-screened samples spanning judgement, single-choice, multiple-choice, short-answer, and open-ended tasks. To test the benchmark's sensitivity to domain adaptation, we used Qwen2.5-7B-Instruct as the base model and applied supervised fine-tuning with LoRA through the LLaMA-Factory framework, yielding GMatLLM. We then evaluated general-purpose, chemistry-specific, materials-specific, and encoder-only models on GlassBench and complementary general benchmarks. GMatLLM achieved approximately 0.72 accuracy on GMat-MCQ and 0.31 on GMat-QA, while its reduced scores on several general benchmarks exposed a domain-specialization/general-capability trade-off. These results indicate that GlassBench can distinguish model families and quantify both the gains and costs associated with domain adaptation. GlassBench therefore provides a reproducible framework for evaluating and developing LLMs for glass materials research.

1 INTRODUCTION

Large language models used in materials research can be broadly grouped by their principal downstream functions into generation-oriented and understanding-oriented systems. Generation-oriented models support tasks such as question answering, scientific reasoning, and materials design. For example, LLaMat and MatterChat adapt language models to materials-specific text and structure-aware interaction (Mishra et al., 2024; Tang et al., 2025). ChatMOF combines an LLM with external tools to retrieve, predict, and generate metal-organic frameworks (MOFs) (Kang & Kim, 2024), whereas CrystaLLM learns an autoregressive representation of crystallographic information files to generate plausible inorganic crystal structures (Antunes et al., 2024). Retrieval-augmented generation provides a complementary strategy by grounding model outputs in external knowledge sources. Understanding-oriented models instead emphasize named-entity recognition, relation extraction, and text classification. Materials-specific encoders such as MatSciBERT and MatBERT have demonstrated the value of domain pretraining for extracting structured information from scientific literature (Gupta et al., 2022; Walker et al., 2022). Together, these approaches illustrate how scientific language models can convert unstructured materials knowledge into searchable, reusable representations.

With the adoption of decoder-only architectures and instruction tuning (Ouyang et al., 2022), contemporary LLMs increasingly perform both understanding and generation tasks. General-purpose model families such as GPT-4, LLaMA, Qwen, Llama 3, and DeepSeek exhibit strong language

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

understanding and generation capabilities (OpenAI, 2023; Touvron et al., 2023; Yang et al., 2024; Grattafiori et al., 2024; DeepSeek-AI et al., 2025), motivating their use as scientific assistants for literature synthesis, hypothesis generation, and experimental planning. Retrieval-augmented generation provides an additional route for grounding answers in traceable external evidence (Lewis et al., 2020). Nevertheless, model performance can be limited in highly specialized domains whose terminology and structure-property relations are underrepresented in general training corpora. Recent studies illustrate the need for domain-aware reasoning in inverse MOF design and catalyst screening (Cleeton & Sarkisov, 2025; Bai et al., 2025). Despite this progress, no widely accepted benchmark systematically evaluates LLM capabilities in individual materials subfields. Evaluations commonly rely on small, study-specific test sets or general benchmarks such as MMLU (Hendrycks et al., 2021), which do not capture the task structure of glass science or directly quantify the gains produced by domain adaptation.

Constructing a domain-specific evaluation system for glass materials presents three challenges. First, major glass databases, including SciGlass and INTERGLAD, primarily contain structured composition-property records rather than natural-language question-answer pairs (Mazurin, 2005; New Glass Forum). Second, the dense terminology and interconnected knowledge modules of glass science make automatically generated questions susceptible to factual errors, semantic ambiguity, and logical discontinuities. Third, broad general-purpose benchmarks provide limited resolution for measuring reasoning in a narrow materials subdomain and may be affected by training-data contamination, which can inflate apparent performance (Sainz et al., 2023). These limitations motivate a controllable and reproducible benchmark designed specifically for glass materials science.

To address this gap, we developed GlassBench, a systematic benchmark for evaluating LLMs in glass materials science. The workflow comprises: (i) extraction of question-answer pairs from authoritative textbooks and research literature; (ii) multistage quality control combining LLM scoring, confident learning, and domain-consistency checks; and (iii) construction of judgement, multiple-choice, short-answer, and open-ended tasks. We then used supervised fine-tuning with low-rank adaptation (LoRA) (Hu et al., 2022) as a validation probe; related quantized adapter methods demonstrate the broader efficiency of parameter-efficient fine-tuning (Dettmers et al., 2023). Fine-tuning is not the central contribution; rather, it tests whether GlassBench can resolve capability differences before and after domain adaptation while also revealing changes in general language performance.

2 METHODS

2.1 BUILDING A DOMAIN-SPECIFIC QUESTION-ANSWERING DATASET

Glass science has accumulated several large composition-property resources. SciGlass and INTERGLAD contain extensive structured records covering diverse glass systems and properties (Mazurin, 2005; New Glass Forum). Such data have supported machine-learning studies of glass-transition temperature, viscosity, dissolution kinetics, and multiple property families (Liu et al., 2021; Cassar et al., 2018; Cassar, 2023; Mauro et al., 2009; Krishnan et al., 2018), while modern definitions emphasize that glass is a nonequilibrium, noncrystalline state whose properties depend on thermal history (Zanotto & Mauro, 2017). These databases are valuable for numerical prediction and composition-property analysis, but their tabular records do not directly provide the natural-language question-answer format needed for LLM evaluation. Their composition distributions, property labels, and process annotations also differ from the knowledge dimensions targeted in GlassBench. We therefore constructed a dedicated evaluation corpus by retrieving domain literature and extracting question-answer pairs, followed by multidimensional quality screening to improve factual accuracy and domain representativeness.

We used a semi-automated workflow in which candidate data were generated by LLMs or rules and then subjected to manual screening and correction. The source corpus comprised authoritative glass-science textbooks (Wang et al., 2006; Wang & Tao, 2008) and research articles. Prior work has shown that materials knowledge, named entities, and synthesis procedures can be extracted at scale from scientific text (Tshitoyan et al., 2019; Weston et al., 2019; Kononova et al., 2019; Swain & Cole, 2016). Bibliographic metadata for more than 3,000 papers was retrieved through the OpenAlex API (Priem et al., 2022) to broaden coverage of emerging glass systems and research directions beyond the selected textbooks.

All source documents underwent standardized preprocessing to normalize file formats and character encoding. MinerU was then used to convert unstructured PDF files into structured Markdown while retaining headings, reading order, tables, and mathematical expressions (Wang et al., 2024). This step reduced common parsing errors such as fragmented text, distorted layouts, and loss of document structure.

The structured documents were processed in Easy Dataset (Miao et al., 2025) using chapter-aware recursive segmentation. Markdown headings defined the initial semantic boundaries; blocks exceeding the prescribed length were further divided at sentence and clause boundaries. In this study, the target block length was 150-300 characters. Each block was linked to source metadata and domain labels before question-answer generation. The initial pipeline produced 30,801 candidate pairs spanning glass fundamentals, raw materials and composition, manufacturing, structure-property relations, characterization, defects, and quality control.

2.2 Data quality filtering

The raw candidate dataset generated during the automatic question-answering extraction stage inevitably contains some quality issues, such as label noise, ambiguous answers, factual errors , or semantic redundancy . To ensure the label quality and inference reliability of the GlassBench evaluation samples, this paper designs a two-stage cascaded cleaning architecture of "coarse screening-fine washing" to systematically audit the quality of the raw candidate question-answer pairs (over 30,000).

2.2.1 Preliminary multidimensional scoring by an LLM

This phase utilizes Qwen-Plus as the judging model (JudgeLLM), using its self-scoring score as a weak label to replace costly manual annotation. Each question-answer pair is scored on a 1-5 scale across three dimensions: factual accuracy, information density, and logical coherence. A hard threshold criterion (score 4) is used for selection: samples are considered "Golden" only if the factual accuracy score is at least 4 and the total score across all three dimensions is at least 12; otherwise, they are moved to the Trash pool. To ensure stability under large-scale concurrent processing, this phase employs multi-threaded concurrent calls (10 worker threads) combined with an exponential backoff retry strategy (maximum wait time 20 seconds, maximum 6 retries) to address API rate limiting. Samples that fail multiple times consecutively are automatically moved to the Trash pool to avoid pipeline congestion, forming a coarse-grained screening mechanism that combines throughput efficiency with fault tolerance.

2.2.2 Statistical screening based on confident learning

Only samples assigned to the Golden pool were subjected to statistical auditing. The factual-correctness score from the first-stage LLM assessment was treated as a weak label. We then constructed two complementary feature sets: 384-dimensional semantic embeddings from all-MiniLM-L6-v2 and term-frequency-inverse-document-frequency features with a maximum vocabulary of 5,000 terms. A logistic-regression classifier with five-fold cross-validation generated out-of-sample predicted probabilities, thereby limiting optimistic bias from in-sample estimates. Cleanlab Datalab was then used to identify potential label issues, outliers, near duplicates, and non-independent-and-identically-distributed samples using the confident-learning framework (Northcutt et al., 2021). Samples flagged as label issues or outliers were removed, and the remaining clean samples were retained for benchmark construction.

2.2.3 Iterative cleaning and convergence

Considering that a single round of statistical detection may not be sufficient to identify all problematic samples, this study adopts an iterative cleaning strategy in which the clean subset identified by Cleanlab in each round is used as the candidate set for the subsequent round. The procedure is repeated until the composition of the retained subset becomes stable, thereby reducing the risk of retaining samples with incorrect labels, anomalous distributions, or semantic redundancy. Importantly, when the candidate sample pool remains unchanged, each cleaning round only removes samples flagged as problematic and therefore cannot increase the sample count. Any apparent increase in sample size would require either the reintroduction of previously removed samples or a

change in the candidate pool, which is not considered part of the standard iterative cleaning procedure adopted here.

As an example, in one intermediate audit round, 542 input samples were examined, Cleanlab identified 83 samples with potentially incorrect labels (15.3%), 20 outliers (3.7%), 2 near-duplicate samples, and 1 non-IID sample. After accounting for these detected issues, 439 samples were retained as the final clean subset, corresponding to an overall removal of 103 problematic samples (19.0%). The iterative procedure therefore serves as a progressive quality-control mechanism rather than a simple one-step filtering operation, with the final subset characterized by more consistent labels, a more reasonable sample distribution, and reduced semantic redundancy.

2.2.4 DOMAIN-KNOWLEDGE CONSISTENCY CHECKS

To further ensure the accuracy of domain-specific knowledge in the dataset, this study conducted a two-step validation process on the filtered dataset. First, a key terminology coverage test was performed , calculating the density of specialized terms in the answers and filtering out answers with excessively low terminology density. Second, cross-validation was conducted. For multiple question-answer pairs related to the same knowledge point, logical contradictions between the answers were checked, and the results were compared with the original text for verification. Finally, 5% of the samples were randomly selected from the automatically filtered dataset and organized by topic for manual review by experts . Experts evaluated the accuracy, completeness, and professionalism of the answers, and manually corrected and deleted problematic samples .

2.3 FINE-TUNING STRATEGY

2.3.1 EXPERIMENTAL DESIGN AND FINE-TUNING PROTOCOL

To assess whether GlassBench distinguishes models with different capability profiles, we used Qwen2.5-7B-Instruct (Yang et al., 2024) as the base model, implemented supervised fine-tuning (SFT) in LLaMA-Factory (Zheng et al., 2024), and adopted LoRA (Hu et al., 2022) as the parameter-efficient fine-tuning strategy. Fine-tuning served as a validation probe for benchmark sensitivity and discriminative power rather than as the primary research contribution.

LoRA's core assumption is that the weight update matrix has a low intrinsic rank when the model is adapted to a specific task. Instead of $W_0 \in \mathrm{R}^{d\times k}$ directly updating the pre-trained weight matrix, LoRA W_0 injects a low-rank decomposition term:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} B \cdot A \tag{2-1}$$

During inference, the final weight is $W = W_0 + \frac{\alpha}{r}\Delta W$, where α is the scaling factor, used to control the influence of the LoRA branch relative to the original weight. During training, W_0 are kept frozen, and only A and B are updated. Among them, $B \in \mathrm{R}^{d\times k}$, $A \in \mathrm{R}^{d\times k}$, and $r \ll \min(d, k)$. In this study, the LoRA rank was set to $r = 16$, and the scaling factor α is taken as 32, yielding $\frac{\alpha}{r} = 2$, and dropout (0.1) was introduced in the LoRA layer to alleviate overfitting. The LoRA modules were applied to all designated linear projection layers. Taking Qwen2.5-7B as the base model, this parameter-efficient configuration freezes the pretrained model parameters and updates only the low-rank adaptation parameters, substantially reducing memory and computational overhead compared with full-parameter fine-tuning while retaining sufficient capacity for domain-specific adaptation.

The training phase is set to standard supervised fine-tuning (stage: SFT). The optimization objective is to x maximize the conditional y log-likelihood of the target answer sequence given the instructions, which is equivalent to minimizing the following cross-entropy loss:

$$\zeta = -\sum_{t} \log P(y_t | y_{<t}, x; \theta) \tag{2-2}$$

In terms of specific training configuration , bfloat16 mixed precision (bf16:true) is used to halve GPU memory usage while maintaining a good numerical dynamic range ; the batch size per GPU is set to 4, and gradient accumulation (gradient_accumulation_steps: 4) achieves an equivalent batch size of 16; the learning rate uses cosine annealing scheduling , with a peak learning rate of 2e-4, and linear warm-up with a 5% warm-up ratio to avoid gradient oscillations in the early stages of training ; the maximum sequence truncation length is set to 2048 tokens; the number of training

epochs is 10 ; and the dialogue format is uniformly encoded using ChatTemplate (Qwen template) that matches the base model . For data partitioning, periodic evaluations are performed every 200 steps with checkpoints saved. After training, the loss curve is automatically plotted to monitor convergence. All hyperparameters and training logs are recorded to ensure the reproducibility of the experiment.Considering that a larger learning rate and a smaller batch size may increase the risk of catastrophic forgetting (Kirkpatrick et al., 2017); The above experiments were conducted on a vGPU server equipped with 32GB of video memory.

2.3.2 External downstream validation on Matbench-glass

To examine whether knowledge acquired through domain-specific fine-tuning transfers beyond the question-answering setting, we evaluated GMatLLM on the independent Matbench-glass task (Dunn et al., 2020). As shown in Table 1, GMatLLM (v0.2) achieves the strongest overall performance among the language-model-based approaches included in the main-text comparison, obtaining a Mean F1 of 0.8808, a Mean ROC-AUC of 0.7806, and a Mean Balanced Accuracy of 0.7806. It slightly outperforms gptchem (Mean F1 = 0.8782; Mean ROC-AUC = 0.7762) and Darwin (Mean F1 = 0.8722; Mean ROC-AUC = 0.7668), indicating that GlassBench is capable of distinguishing between closely performing domain-specific language models.

A clear performance gap is observed between the stronger and weaker models. ChemDFM-v1.5-8B achieves a Mean F1 of 0.7792 and a Mean ROC-AUC of 0.5397, while GMatLLM (v0.1) obtains 0.7589 and 0.5469, respectively. Notably, GMatLLM (v0.2) substantially improves over its earlier version, with increases of 0.1219 in Mean F1 and 0.2337 in Mean ROC-AUC, demonstrating that the benchmark is sensitive to changes in model capability and domain adaptation.

The Dummy baseline yields a Mean ROC-AUC and Mean Balanced Accuracy of approximately 0.50, consistent with chance-level performance. Together, these results demonstrate that GlassBench provides a sufficiently broad and graded performance landscape to distinguish models across different capability levels, while retaining adequate resolution among closely performing glass-domain language models.

Table 1: Downstream performance on Matbench-glass.

Model	Mean F1	Mean ROCAUC	Mean Balanced Accuracy	Std Rocauc
GMatLLM(v0.2)	0.8808	0.7806	0.7806	0.0282
gptchem	0.8782	0.7762	0.7762	0.0122
Darwin	0.8722	0.7668	0.7668	0.0133
ChemDFM-v1.5-8B	0.7792	0.5397	0.5397	0.0099
GMatLLM(v0.1)	0.7589	0.5469	0.5469	0.0169
Dummy baseline	0.7127	0.5005	0.5005	0.0178

2.4 Benchmark design

To verify GlassBench's ability to distinguish between model capabilities, the benchmark design needs to meet two conditions simultaneously: First, the models participating in the comparison should cover different training backgrounds (general vs. domain, before fine-tuning vs. after fine-tuning) to observe whether GlassBench can capture the capability differences corresponding to these background differences; second, the test set participating in the comparison should include both general capability and domain capability dimensions to place GlassBench's domain evaluation results in a comparable coordinate system, rather than viewing a single score in isolation. Based on this, this paper selects the evaluation model and test set as follows.

(1) Evaluation-model selection. We selected a representative set of models spanning three categories, totaling 18. General-purpose LLMs included GPT-4.1, Qwen2.5, LLaMA-3-8B, and DeepSeek-R1-Distill-Qwen-7B (OpenAI, 2023; Yang et al., 2024; Grattafiori et al., 2024; DeepSeek-AI et al., 2025). Chemistry- and materials-oriented models, including ChemLLM, ChemDFM, and BatteryBERT, were included to test whether adaptation to an adjacent scientific domain transfers to glass science (Zhang et al., 2024; Zhao et al., 2024; Huang & Cole, 2022). Finally, GMatLLM, obtained by LoRA-based supervised fine-tuning of Qwen2.5-7B-Instruct on the GlassBench training set, was compared directly with its unfine-tuned base model. The design

therefore covers open- and closed-source systems, general- and domain-specific models, and before-versus-after-fine-tuning comparisons.

(2) Benchmark-dataset selection. The general-capability benchmarks comprised MMLU (Hendrycks et al., 2021), CMMLU (Li et al., 2023), TruthfulQA (Lin et al., 2022), AIME-2025, and GPQA-Diamond (Rein et al., 2023). Together, these benchmarks assess broad knowledge, factuality, mathematical reasoning, and graduate-level scientific reasoning. The domain-specific evaluation used the GlassBench test set, which covers glass fundamentals, composition design, processing, structure-property relations, and characterization. We report GMat-MCQ for judgement and multiple-choice items and GMat-QA for short-answer and open-ended items, preventing a single aggregate score from obscuring differences between structured selection and free-form generation.

Metrics were selected according to task format. Accuracy was used for the multiple-choice benchmarks after extracting the model's selected option with regular expressions. For AIME-2025, the final answer was parsed and evaluated by numerical equivalence. GMat-MCQ was scored by accuracy, whereas GMat-QA combined BERTScore (Zhang et al., 2020) with an LLM-as-judge assessment to accommodate semantically equivalent responses expressed in different wording. Because LLM judges can exhibit position, verbosity, and self-enhancement biases, their configuration and validation should be reported explicitly (Zheng et al., 2023). All reported metrics were computed within the EvalScope framework (ModelScope Team) to promote consistent model and dataset handling. The overall process of the project is shown in Figure 1.

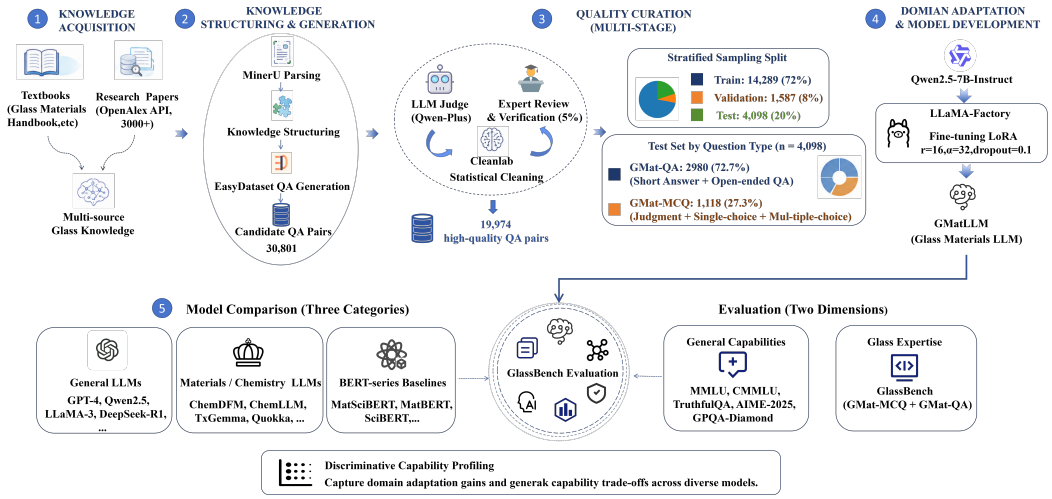

Figure 1: Construction and validation framework of GlassBench. the workflow includes knowledge acquisition, document structuring and question-answer generation, multistage quality control, benchmark construction and domain adaptation, and comparative evaluation across general-purpose and domain-specific models.

3 Results and Discussion

3.1 Final Dataset Statistics

After multistage screening and cleaning, the final GlassBench dataset contained 19,974 question-answer pairs across five task categories: judgement, single choice, multiple choice, short answer, and open answer. These categories form a progression from knowledge recognition and selection to concise factual retrieval and open-ended domain-specific generation.

GlassBench contains 14,289 training samples, 4,098 test samples, and 1,587 validation samples, corresponding to 71.5%, 20.5%, and 7.9% of the full dataset, respectively. All three datasets were constructed using stratified sampling to ensure reasonable domain coverage for each subset,as shown in Fig. 2. The dataset is thematically concentrated. The glass science domain dominates the training

set, with 13,435 records (94.0%), representing the vast majority. These domains, arranged in descending order of quantity, are: other glass domains (731 records), physics (481 records), etc., forming a "single-core, multi-wing" distribution pattern, fully reflecting the specialized vertical focus of the dataset. The domain distribution of the test set differs significantly from that of the training set, and is more balanced and diverse. The glass science and engineering field accounts for 64.3% (2634 items), while other glass categories account for 875 items. The difference in distribution between the training and test sets is intentional—the test set has a higher proportion of cross-domain samples and comprehensive questions, which helps to more comprehensively evaluate the knowledge transfer ability of the fine-tuned model and its robustness outside the core domain, thus avoiding overestimation bias in the evaluation results caused by overly concentrated distribution.

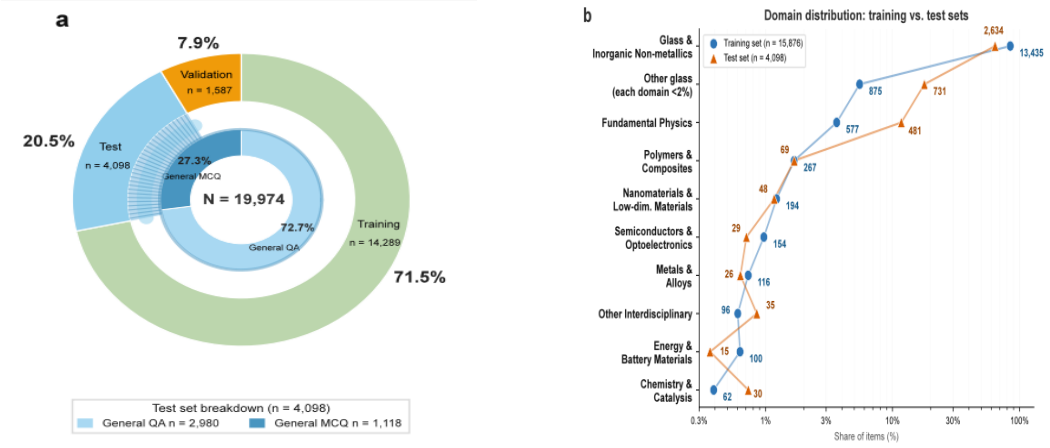

Figure 2: Dataset composition and domain distribution. (a) train/test/validation proportions for the full dataset (N = 19,974) and the composition of the held-out test set. (b) domain distributions of the training and test sets.

3.2 Benchmark Performance

GlassBench produced a broad, continuous distribution of model scores. On GMat-MCQ (Figures 3 and 4), scores ranged from 0.10 for LlaSMol-Mistral-7B to approximately 0.74 for the highest-performing models. GMat-QA scores ranged from 0.01 for Quokka-7B-Instruct to 0.31 for GMatLLM. The dispersed score distributions indicate that the benchmark avoids strong ceiling and floor effects and can resolve capability differences among models with distinct architectures, scales, and training histories.

The comparative analysis of general-purpose large language models clearly reveals the evolutionary pattern of their capabilities in the glass materials field, confirming that there is no simple linear positive correlation between model domain adaptability and parameter size. Simply expanding model size cannot effectively improve the professional question-answering capabilities in specific domains. The evaluation data shows that mainstream large-parameter general-purpose models have significant advantages in structured selection tasks. Among them, gpt-4.1 and Qwen2.5-32B-Instruct performed exceptionally well in the GMat-MCQ task, ranking among the top general-purpose models, demonstrating the fundamental advantages of large-scale general-purpose models in basic domain knowledge recognition and single-choice discrimination tasks. However, in the GMat-QA open-ended question-answering task, which is more closely aligned with practical application scenarios and tests deep domain understanding and autonomous answering capabilities, the performance shortcomings of the two large-parameter models were fully exposed, with scores of only 0.20 and 0.26 respectively. Their performance was significantly inferior to GMatLLM (0.30), which has a much smaller parameter size and has been specifically adapted for the glass materials field. Meanwhile, a horizontal comparison with other gradient-scale models in the same series reveals that Qwen2.5-14B, with a moderate number of parameters, achieved a score of 0.26 in the GMat-QA task. This score is only slightly higher than the 0.24 score of the Qwen2.5-7B base model, which has even fewer parameters, indicating that performance did not show a significant leap with increasing pa-

rameter count. These comparative results clearly demonstrate that blindly expanding the parameter count of a general-purpose language model has extremely limited effect on improving the open-ended, generative question-answering capabilities in the glass materials field. It cannot compensate for the core shortcoming of general-purpose models lacking domain-specific knowledge adaptation. Targeted domain fine-tuning and knowledge adaptation optimization are far more suitable for the task requirements of the glass materials subfield than simply increasing the model size. This is the key reason why this paper abandons relying on a general-purpose model with extremely large parameter counts as the experimental baseline and chooses the GMatLLM obtained from the lightweight model adaptation as the core validation case.

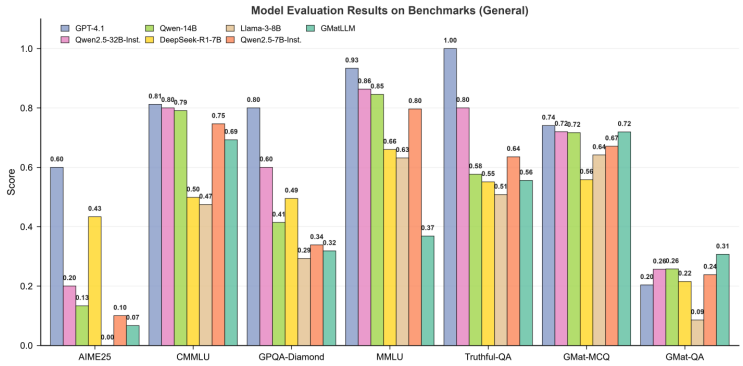

Figure 3: Performance of general-purpose large language models across general and domain-relevant benchmarks. the grouped bars compare seven models on AIME-2025, CMMLU, MMLU, GPQA-Diamond, TruthfulQA, General MCQ, and General QA. scores are reported on a 0-1 scale. the figure illustrates task-dependent differences among model families and highlights the difficulty of open-ended scientific question answering.

Comparisons across chemistry-, materials-, and glass-specific models further demonstrate the domain specificity of GlassBench. The chemistry-oriented models (Zhang et al., 2024; Zhao et al., 2024) and the battery-domain encoder (Huang & Cole, 2022) did not consistently transfer to glass-specific tasks: even the strongest adjacent-domain baseline remained below the glass-adapted GMatLLM on the core domain evaluations. Materials-oriented generative models also showed limited performance. These findings suggest that broad chemistry or materials pretraining does not automatically provide the formulation, processing, property, and engineering knowledge required in glass science, thereby supporting the need for a dedicated vertical-domain benchmark.

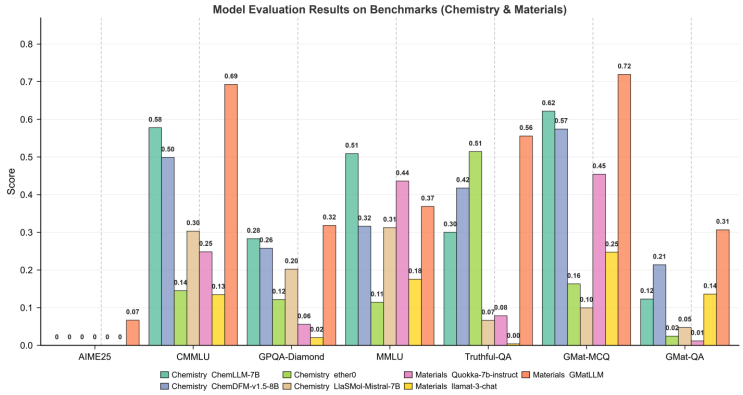

Figure 4: Performance of chemistry- and materials-specialized models across general and domain-relevant benchmarks. the grouped bars compare chemistry-specific and materials-oriented models on seven tasks. the results show narrow, task-dependent competence profiles and limited transfer from adjacent scientific domains to glass materials science.

3.3 ENCODER-ONLY BASELINES

To provide cross-architecture context, we also evaluated encoder-only models, including BERT-base, SciBERT, MatBERT, and MatSciBERT (Gupta et al., 2022; Walker et al., 2022; Beltagy et al., 2019; Devlin et al., 2019). These models fundamentally differ from instruction-tuned decoder-only LLMs in architectural design, pretraining objectives, and inference paradigms. Accordingly, we adopt tailored evaluation protocols for each model type instead of a unified zero-shot generation setting. Each question is paired with individual candidate options for model matching and scoring, and the highest-scoring option is selected as the prediction. Training and independent test sets contain 2,362 and 223 samples, respectively. In contrast, decoder-only models (e.g., GMatLLM) are evaluated via zero-shot conditional log-probability scoring without task-specific fine-tuning.

This paradigm-aligned evaluation conforms to the standard usage of each model type and prevents unfairly attributing the inherent absence of autoregressive generation in encoder-only models to insufficient domain knowledge. All encoder-only models outperform the 25% random baseline on GMat-MCQ but lag behind GMatLLM. For instance, MatSciBERT achieves 40.8% accuracy, while GMatLLM reaches 73.8% (Figures 5). Given the inherent differences in model scale, pretraining data, optimization objectives, supervision settings, and inference pipelines, this performance gap serves only as a cross-paradigm reference rather than a strictly controlled architectural comparison.

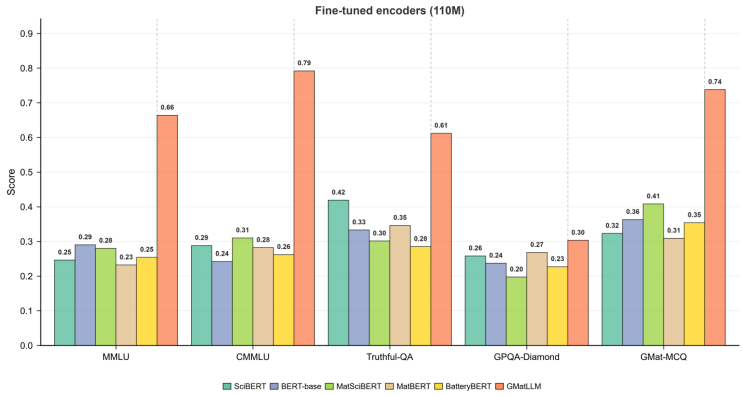

Figure 5: Benchmark performance of domain-specific masked language models. five encoder-only models are compared with GMatLLM across general and domain-relevant tasks. the results provide a cross-paradigm reference; they should not be interpreted as a controlled comparison because the systems differ in architecture, supervision, and evaluation protocol.

4 CONCLUSIONS

This study introduces GlassBench, a multitask benchmark for evaluating LLMs in glass materials science. The benchmark combines domain-specific question-answer construction, multistage quality control, and complementary structured and open-ended evaluation tasks. Experiments across general-purpose, chemistry-specific, materials-specific, glass-adapted, and encoder-only models produced a broad distribution of scores, demonstrating the benchmark's ability to distinguish capability profiles. Using supervised LoRA fine-tuning as a validation probe, GlassBench captured gains in glass-domain performance as well as losses on several general benchmarks. These findings show why domain evaluation should report both specialized capability and general-capability retention. Cross-architecture results are included as contextual baselines rather than strict controlled comparisons.

Future work will extend GlassBench to multimodal inputs, including glass-structure images and spectroscopic data, and will evaluate retrieval-augmented generation, tool use, reinforcement learning, and less aggressive adaptation strategies. Further releases should also provide complete data provenance, deduplication and contamination analyses (Sainz et al., 2023), expert-agreement statistics, and fixed evaluation scripts to strengthen reproducibility and support fair comparison across models.

AI USE STATEMENT

In this work, generative-AI tools were used for manuscript language polishing, sentence rephrasing, and grammatical refinement, which falls under recommended-disclosure tasks. No generative-AI tools were used for required-disclosure tasks including research ideation, dataset construction, experimental design, deriving scientific arguments, or analyzing experimental results.

All AI-assisted text was fully reviewed, revised and verified by the authors. The authors take full responsibility for all content of this work, including all text, claims, and artifacts

REFERENCES

L. M. Antunes, K. T. Butler, and R. Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15:10570, 2024. doi: 10.1038/s41467-024-54639-7.

R. Bai, Y. Yao, Q. Lin, et al. Preferable single-atom catalysts enabled by natural language processing for high-energy-density Na-S batteries. *Nature Communications*, 16:5827, 2025. doi: 10.1038/s41467-025-60931-x.

I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615–3620, 2019. doi: 10.18653/v1/D19-1371.

D. R. Cassar. GlassNet: A multitask deep neural network for predicting many glass properties. *Ceramics International*, 49(22B):36013–36024, 2023. doi: 10.1016/j.ceramint.2023.08.281.

D. R. Cassar, A. C. P. L. F. de Carvalho, and E. D. Zanotto. Predicting glass transition temperatures using neural networks. *Acta Materialia*, 159:249–256, 2018. doi: 10.1016/j.actamat.2018.08.022.

C. Cleeton and L. Sarkisov. Inverse design of metal-organic frameworks using deep dreaming approaches. *Nature Communications*, 16:4806, 2025. doi: 10.1038/s41467-025-59952-3.

DeepSeek-AI et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

T. Dettmers, A. Pagnoni, A. Holtzman, et al. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023.

J. Devlin, M. W. Chang, K. Lee, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

A. Dunn, Q. Wang, A. Ganose, et al. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6:138, 2020. doi: 10.1038/s41524-020-00406-3.

A. Grattafiori, A. Dubey, A. Jauhri, et al. The Llama 3 herd of models. *arXiv [Preprint]*, 2024. Available from: `https://arxiv.org/abs/2407.21783`.

T. Gupta, M. Zaki, N. M. A. Krishnan, et al. MatSciBERT: A materials-domain language model for text mining and information extraction. *npj Computational Materials*, 8:102, 2022. doi: 10.1038/s41524-022-00784-w.

D. Hendrycks, C. Burns, S. Basart, et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

E. J. Hu, Y. Shen, P. Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

S. Huang and J. M. Cole. BatteryBERT: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling*, 62(24):6365–6377, 2022. doi: 10.1021/acs.jcim.2c00035.

Y. Kang and J. Kim. ChatMOF: An artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature Communications*, 15:4705, 2024. doi: 10.1038/s41467-024-48998-4.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 (13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

O. Kononova, H. Huo, T. He, et al. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6:203, 2019. doi: 10.1038/s41597-019-0224-1.

N. M. A. Krishnan, S. Mangalathu, M. M. Smedskjaer, et al. Predicting the dissolution kinetics of silicate glasses using machine learning. *Journal of Non-Crystalline Solids*, 487:37–45, 2018. doi: 10.1016/j.jnoncrysol.2018.02.023.

P. Lewis, E. Perez, A. Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

H. Li, Y. Zhang, F. Koto, et al. CMMLU: Measuring massive multitask language understanding in Chinese. *arXiv [Preprint]*, 2023. Available from: https://arxiv.org/abs/2306.09212.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

H. Liu, Z. Fu, K. Yang, et al. Machine learning for glass science and engineering: A review. *Journal of Non-Crystalline Solids*, 557:119419, 2021. doi: 10.1016/j.jnoncrysol.2019.04.039.

J. C. Mauro, Y. Yue, A. J. Ellison, et al. Viscosity of glass-forming liquids. *Proceedings of the National Academy of Sciences of the United States of America*, 106(47):19780–19784, 2009. doi: 10.1073/pnas.0911705106.

O. V. Mazurin. Glass properties: Compilation, evaluation, and prediction. *Journal of Non-Crystalline Solids*, 351:1103–1112, 2005. doi: 10.1016/j.jnoncrysol.2005.01.024.

Z. Miao, Q. Sun, J. Wang, 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*, pp. 960–968, 2025.

V. Mishra, S. Singh, D. Ahlawat, et al. LLaMat: Large language models for materials science information extraction. In *NeurIPS Workshop*, 2024.

ModelScope Team. EvalScope: A framework for large-model evaluation. Available from: https://github.com/modelscope/evalscope. Accessed: 2026-08-24.

New Glass Forum. INTERGLAD: International glass database system. Available from: https://www.interglad.jp/apps/. Accessed: 2026-08-24.

C. G. Northcutt, L. Jiang, and I. L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: 10.1613/jair.1.12125.

OpenAI. GPT-4 technical report. *arXiv [Preprint]*, 2023. Available from: https://arxiv.org/abs/2303.08774.

L. Ouyang, J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

J. Priem, H. Piwowar, and R. Orr. OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. *arXiv [Preprint]*, 2022. Available from: https://arxiv.org/abs/2205.01833.

D. Rein, B. L. Hou, A. C. Stickland, et al. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv [Preprint]*, 2023. Available from: https://arxiv.org/abs/2311.12022.

O. Sainz, J. Campos, I. Garcia-Ferrero, et al. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

M. C. Swain and J. M. Cole. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10): 1894–1904, 2016. doi: 10.1021/acs.jcim.6b00207.

Y. Tang, W. Xu, J. Cao, et al. MatterChat: A multi-modal LLM for material science. *arXiv [Preprint]*, 2025. Available from: https://arxiv.org/abs/2502.13107.

H. Touvron, T. Lavril, G. Izacard, et al. LLaMA: Open and efficient foundation language models. *arXiv [Preprint]*, 2023. Available from: https://arxiv.org/abs/2302.13971.

V. Tshitoyan, J. Dagdelen, L. Weston, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95–98, 2019. doi: 10.1038/s41586-019-1335-8.

N. Walker, A. Trewartha, H. Huo, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3:100488, 2022. doi: 10.1016/j.patter.2022.100488.

B. Wang, C. Xu, X. Zhao, et al. MinerU: An open-source solution for precise document content extraction. *arXiv [Preprint]*, 2024. Available from: https://arxiv.org/abs/2409.18839.

Chengyu Wang and Ying Tao. *Handbook of Glass Materials*. Chemical Industry Press, Beijing Shi, beijing di 1 ban edition, 2008.

Chengyu Wang, Min Chen, and Jianhua Chen. *Glass Manufacturing Processes*. Chemical Industry Press, Beijing Shi, di 1 ban edition, 2006.

L. Weston, V. Tshitoyan, J. Dagdelen, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702, 2019. doi: 10.1021/acs.jcim.9b00470.

A. Yang, B. Yang, B. Hui, et al. Qwen2 technical report. *arXiv [Preprint]*, 2024. Available from: https://arxiv.org/abs/2407.10671.

E. D. Zanotto and J. C. Mauro. The glassy state of matter: Its definition and ultimate fate. *Journal of Non-Crystalline Solids*, 471:490–495, 2017. doi: 10.1016/j.jnoncrysol.2017.05.019.

D. Zhang, W. Liu, Q. Tan, et al. ChemLLM: A chemical large language model. *arXiv [Preprint]*, 2024. Available from: https://arxiv.org/abs/2402.06852.

T. Zhang, V. Kishore, F. Wu, et al. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Z. Zhao, D. Ma, L. Chen, et al. ChemDFM: A large language foundation model for chemistry. *arXiv [Preprint]*, 2024. Available from: https://arxiv.org/abs/2401.14818.

L. Zheng, W. L. Chiang, Y. Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Y. Zheng, R. Zhang, J. Zhang, et al. LlamaFactory: Unified efficient fine-tuning of 100+ language models. *arXiv [Preprint]*, 2024. Available from: https://arxiv.org/abs/2403.13372.

A APPENDICES

A.1 TWO-STAGE CASCADED CLEANING

After the two-stage cascaded cleaning and multiple rounds of iterative convergence, the evaluation samples finally retained by GlassBench have achieved a high level in terms of label confidence, semantic independence and distribution consistency, as shown in Figure 6, providing a solid data quality guarantee for the discrimination and reproducibility of subsequent benchmarks.

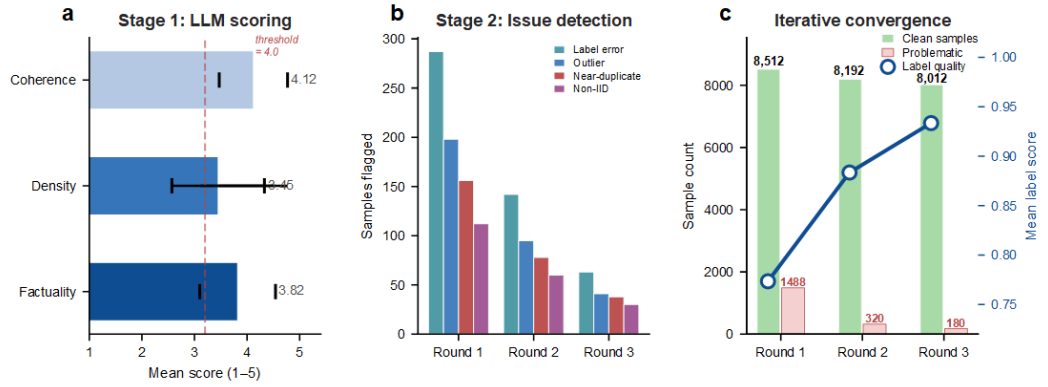

Figure 6: Quantitative evaluation of the two-stage data-cleaning pipeline. (a) LLM-based quality scores across three dimensions (mean $\pm$ standard deviation); the dashed line indicates the screening threshold. (b) Issues count detected by CleanLab across three iterations. (c) Taking ten thousand as an example, retained and removed sample counts and the corresponding mean label-quality score across iterations.

A.2 MULTI-DIMENSIONAL KNOWLEDGE DISTRIBUTION IN GLASS SCIENCE QA SAMPLES

Furthermore, to illustrate the heterogeneous knowledge required in glass materials science, Figure 7 presents four representative question-answer pairs from GlassBench, with answer components color-coded by semantic role: orange for domain terminology and concepts, green for engineering insights and practical considerations, and purple for material properties, performance characteristics, and advantages.

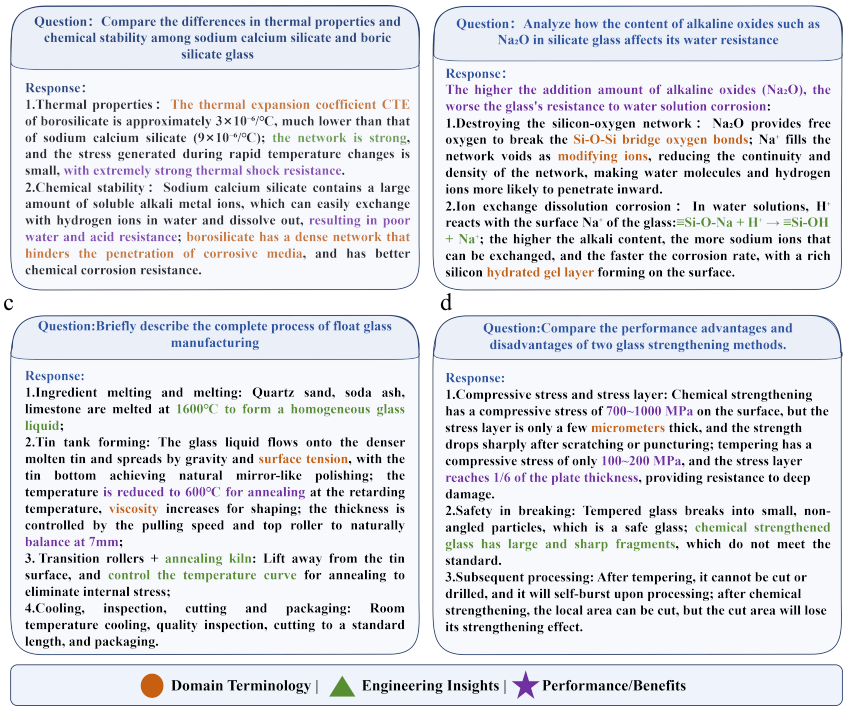

Figure 7: Multidimensional knowledge representation in GlassBench. Four representative question-answer pairs illustrate the heterogeneous knowledge required in glass materials science. Answer components are color-coded by semantic role: orange (domain terminology and concepts), green (engineering insights and practical considerations), and purple (material properties, performance characteristics, and advantages).

A.3 Supplementary Baselines for Conventional Materials ML Models

For completeness, several high-performing conventional materials machine-learning models, including MODNet (v0.1.12), RF-SCM/Magpie, MODNet (v0.1.10), and AMMExpress v2020, are reported separately in Table 2 in the Supplementary Information. These models achieve higher numerical scores on this particular evaluation, but they belong to a different model family and are therefore not used as direct competitors in the main-text LLM comparison. This separation allows the main analysis to focus on the relative capabilities of language-model-based systems while preserving the complete external benchmark results for reference.

Table 2: Downstream performance on Matbench-glass.

Model	Mean F1	Mean ROCAUC	Mean Balanced Accuracy	Std Rocauc
MODNet (v0.1.12)	0.9784	0.9603	0.9603	0.0075
RF-SCM / Magpie	0.9278	0.8587	0.8587	0.0158
MODNet (v0.1.10)	0.9104	0.8107	0.8107	0.0212
AMMExpress v2020	0.9043	0.8607	0.8607	0.0199
GMatLLM(v0.2)	0.8808	0.7806	0.7806	0.0282
gptchem	0.8782	0.7762	0.7762	0.0122
Darwin	0.8722	0.7668	0.7668	0.0133
ChemDFM-v1.5-8B	0.7792	0.5397	0.5397	0.0099
GMatLLM(v0.1)	0.7589	0.5469	0.5469	0.0169
Dummy baseline	0.7127	0.5005	0.5005	0.0178

A.4 CATEGORY-WISE REPRESENTATIVE EXAMPLES OF KNOWLEDGE COMPONENTS

Figure 8 presents one representative example from each category and highlights the knowledge components required in the corresponding answers.

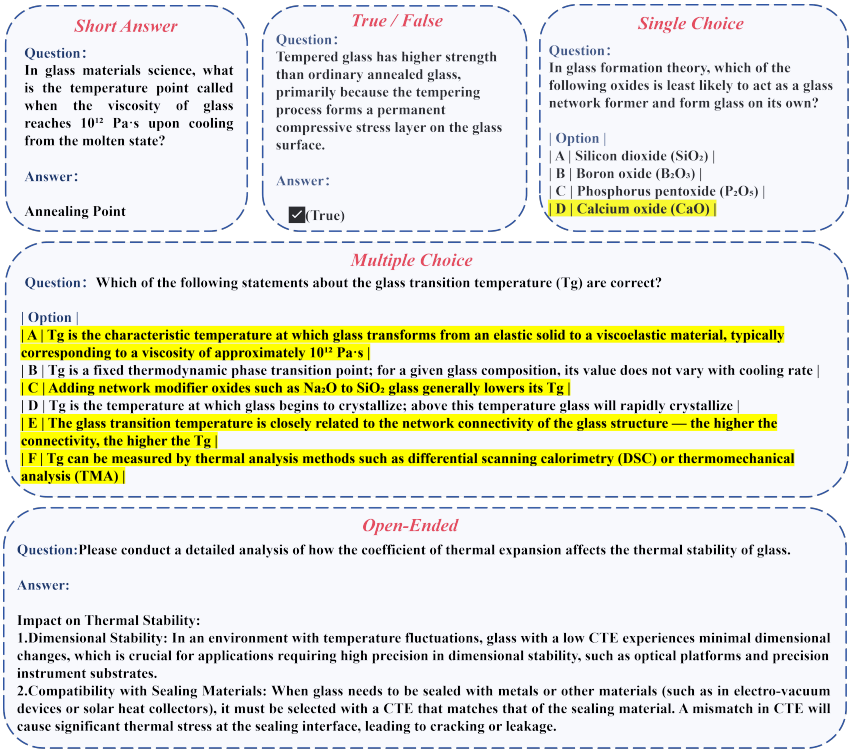

Figure 8: Representative examples of the five GlassBench task types: judgement, single choice, multiple choice, short answer, and open answer. color highlights identify domain terminology, engineering considerations, and material properties or performance information.

A.5 QUALITATIVE COMPARISON OF MODEL ANSWER CONTENTS

However, it is still difficult to determine whether the performance differences between different models correspond to real differences in domain knowledge coverage based solely on the distribution of the final scores. To further verify that the model capability differences characterized by GlassBench have a clear knowledge basis, this paper further conducts a qualitative comparison of different types of models at the level of answer content.

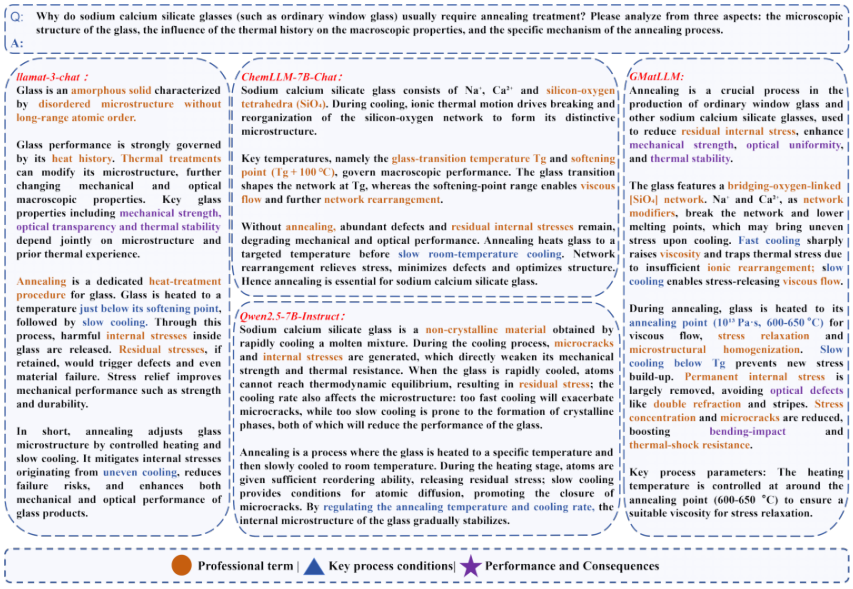

Figure 9: Qualitative comparison of domain-knowledge coverage across model families for the same glass-materials question. responses are annotated for professional terminology, key process conditions, and material-performance consequences. the comparison illustrates differences in coverage and organization; it is complementary to, rather than a substitute for, quantitative evaluation. answer components are color-coded by semantic role: orange indicates professional term, blue indicates key process conditions, and purple indicates performance and consequences.

As shown in Figure 9, the evaluated model families exhibited distinct knowledge-coverage patterns for the same glass-science question. General-purpose models supplied basic facts but integrated glass terminology, engineering practice, and property relationships less completely. Chemistry- and materials-specific models covered more specialized concepts but still omitted some glass-specific context. GMatLLM produced the most integrated response across the annotated knowledge dimensions.

Relative to Qwen2.5-7B-Instruct, GMatLLM improved from 0.67 to 0.72 on GMat-MCQ (an increase of 5 percentage points) and from 0.24 to 0.31 on GMat-QA (an increase of 7.2 percentage points). At the same time, performance decreased on several general benchmarks: MMLU fell from 0.80 to 0.37, while CMMLU, TruthfulQA, GPQA-Diamond, and AIME-2025 also declined. This pattern indicates a trade-off between domain specialization and general capability. The aggressive fine-tuning configuration may have contributed to catastrophic forgetting; however, this causal interpretation should be confirmed through controlled ablations of learning rate, LoRA rank, target modules, training duration, and replay or regularization strat egies.These results show that GlassBench captures both specialized gains and accompanying changes in general capability. Figure 10 therefore analyzes GMatLLM at three levels.

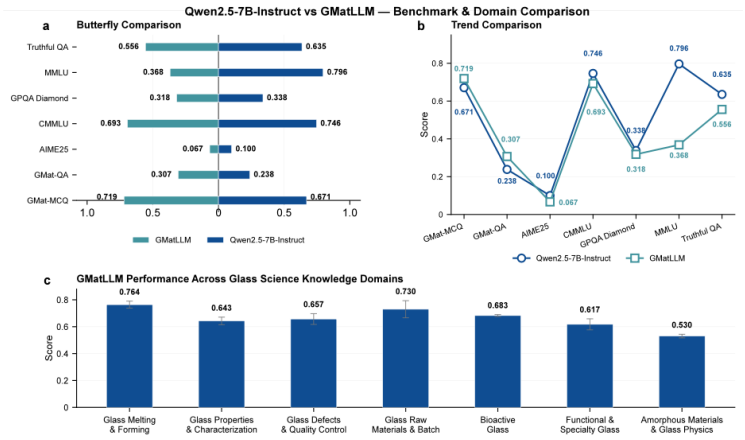

Figure 10: General-benchmark performance and domain-specific capabilities of GMatLLM. (a) comparison of GMatLLM and Qwen2.5-7B-Instruct across general and domain-relevant benchmarks. (b) task-level score profiles. (c) GMatLLM performance across seven glass-science knowledge domains.