

Benchmarking for Domain-Specific LLMs: A Case Study on Academia and Beyond

Rubing Chen¹, Jiaxin Wu¹, Jian Wang¹, Xulu Zhang¹,
Wenqi Fan¹, Chenghua Lin², Xiao-Yong Wei^{1,*}, Qing Li¹

¹The Hong Kong Polytechnic University ²The University of Manchester

Abstract

The increasing demand for domain-specific evaluation of large language models (LLMs) has led to the development of numerous benchmarks. These efforts often adhere to the principle of data scaling, relying on large corpora or extensive question-answer (QA) sets to ensure broad coverage. However, the impact of corpus and QA set design on the *precision* and *recall* of domain-specific LLM performance remains poorly understood. In this paper, we argue that data scaling is not always the optimal principle for domain-specific benchmark construction. Instead, we introduce COMP-COMP, an iterative benchmarking framework grounded in the principle of *comprehensiveness* and *compactness*. Comprehensiveness ensures semantic recall by covering the full breadth of the domain, while compactness improves precision by reducing redundancy and noise. To demonstrate the effectiveness of our approach, we present a case study conducted at a well-renowned university, resulting in the creation of PolyBench, a large-scale, high-quality academic benchmark. Although this study focuses on academia, the COMP-COMP framework is domain-agnostic and readily adaptable to a wide range of specialized fields. The source code and datasets can be accessed at <https://github.com/AnyarB-Chen/COMP-COMP>.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide range of tasks (OpenAI, 2024). While general-purpose LLMs excel in broad applications, domain-specific LLMs have emerged as critical tools for delivering precise and accurate responses tailored to specific user needs (Fei et al., 2023; Cai et al., 2024; Chang et al., 2024). These models are particularly valuable in fields where fixed-answer questions are prevalent,

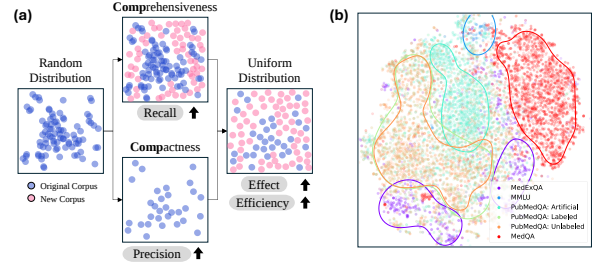


Figure 1: Illustration of *comprehensiveness* and *compactness* in benchmarking. (a) Comprehensiveness ensures broad semantic coverage of the domain, while compactness improves semantic precision by minimizing redundancy, leading to a more uniform data distribution in domain corpora. (b) Semantic distributions of representative benchmarks in the medical domain.

such as law, medicine, and academic, where precision is paramount (Ge et al., 2024). However, the development of domain-specific LLMs necessitates high-quality benchmarks that can effectively evaluate their performance. Such benchmarks must not only ensure accurate responses within the domain, but also maintain broad coverage of the domain’s semantic space to avoid catastrophic forgetting, where models lose proficiency in broader contexts (Luo et al., 2023). This dual requirement underscores the need to develop effective approaches for domain-specific benchmarking.

Previous studies on constructing domain-specific benchmarks have primarily relied on the principle of data scaling, utilizing extensive corpora or large question sets to ensure broad coverage (Cai et al., 2024; Fei et al., 2023). While this approach is widespread and forms the basis for many benchmark designs, it presents several limitations that can affect the efficacy of model evaluation. First, many benchmarks are derived from skill examinations, practice exercises, or expert-curated datasets, often overlooking a critical assessment of their effectiveness (Liu et al., 2023; Ling et al., 2023). Although these benchmarks are diverse in question-

*Corresponding author

Benchmark	Domain	Closeness	QA Type	# Questions
SciEval (Cai et al., 2024)	Science	Open	MCQ, QA	16,522
DISC-Law-Eval (Yue et al., 2023)	Law	Open	MCQ, MAQ, QA	2,863
Counseling Bench (Liu et al., 2023)	Medical	Open	QA	229
CMtMedQA (Yang et al., 2024)	Medical	Open	QA	70,000
COVID-QA (Möller et al., 2020)	Healthcare	Closed	QA	2,019
WC2014QA (Zhang et al., 2016)	Sports	Closed	QA	10,211
PolyBench (Ours)	Academic	Closed	Binary, MAQ, QA, MCQ	24,883

Table 1: Comparison between domain-specific benchmarks.

answering types, ranging from multiple-choice to open-ended formats, they fail to systematically examine the semantic alignment between questions and the underlying corpus (Yue et al., 2023; Cui et al., 2023). This lack of alignment means that the benchmarks may not fully represent the domain’s knowledge structure, leading to evaluations that do not accurately reflect a model’s true capabilities within the specific domain. Second, domain-specific benchmarks often employ LLMs for scoring but neglect to ensure consistency between training and evaluation datasets (Zhang et al., 2023; Yang et al., 2024). This oversight can introduce biases where models inadvertently perform better on evaluation tasks due to similarities with the training data, rather than a genuine understanding of the domain. It highlights the need for a more balanced approach to benchmarking, one that integrates both horizontal coverage across a wide range of topics and vertical depth within specific subdomains.

In this paper, we propose **Comprehensiveness-Compactness (COMP-COMP)**, a novel framework for constructing domain-specific benchmarks that dynamically balances the semantic distribution of corpora and benchmark questions. As illustrated in Figure 1, COMP-COMP iteratively expands the range of corpora and questions, continuously assessing their comprehensiveness (maximizing recall) and compactness (optimizing precision). The framework introduces three crucial mechanisms: (1) *Comprehensiveness and Compactness Monitoring*: We encode the corpora and questions into a unified semantic space, using Gaussian Kernel Density Estimation (KDE) to improve these metrics iteratively (Terrell and Scott, 1992). (2) *Corpora-Question Interaction*: We ensure comprehensive evaluation by generating new questions from under-represented areas in the corpora. (3) *Superiority in Closed Domains*: We prioritize achieving a balance between comprehensiveness and compactness, ensuring the framework excels in constructing closed-

domain benchmarks that demand precise corpus evaluation and accurate answer verification.

To validate our framework, we conducted a case study in a well-renowned university and applied COMP-COMP to develop **PolyBench**, one of the first and most extensive benchmarks for the academic domain (see Table 1). PolyBench includes 24.9k questions across various formats and incorporates user-interest-oriented forum discussions, setting a new standard for domain-specific benchmarking. Extensive experiments on the PolyBench demonstrate the practicality and effectiveness of COMP-COMP, providing a scalable approach for future benchmarking in other domains.

In summary, our contributions are as follows: (1) We propose COMP-COMP, a principled approach that balances comprehensiveness and compactness in domain-specific benchmark construction. (2) We develop PolyBench, a large-scale, high-quality benchmark for the academic domain, which addresses the gap of lacking well-established academic LLM benchmarks. (3) Our method is extensible across different domains, offering valuable insights for future benchmarking. By addressing the limitations of existing methods and providing a systematic benchmark design, our work advances the development of domain-specific LLMs.

2 Related Work

Domain-Specific Benchmarks Domain-specific evaluation has seen significant advances across critical verticals, each adopting distinct strategies reflective of their unique demands. In the legal domain, systems like DISC-LawLLM (Yue et al., 2023) employ a dual evaluation paradigm: standardized law examinations (MCQs with tiered difficulty) paired with ChatGPT, arbitrated subjective assessments. LawBench (Fei et al., 2023) further refines this by grounding evaluations in Bloom’s Taxonomy (Krathwohl, 2002), systematically testing knowledge memorization, understanding, and

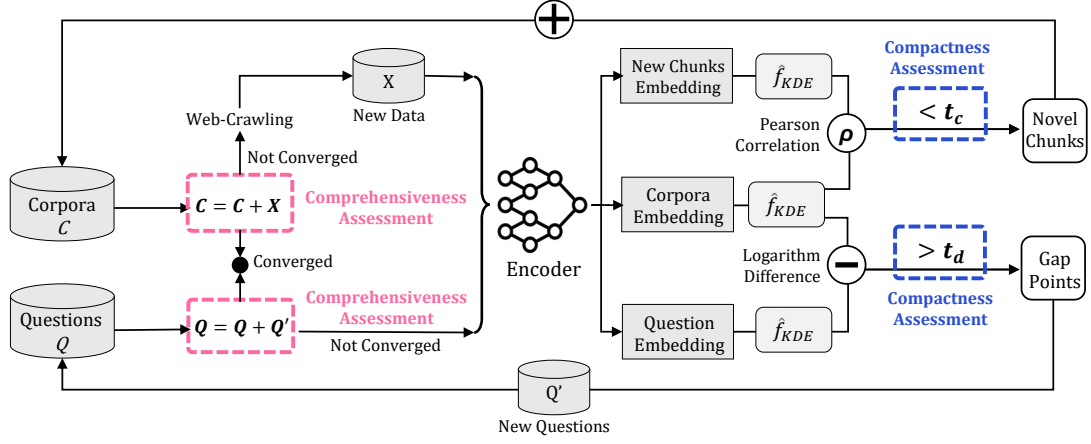


Figure 2: Overview of the proposed Comprehensiveness-Compactness (COMP-COMP) framework for domain-specific benchmark construction. By imposing restrictions represented by t_c and t_d , we aim to enhance precision through compactness. Meanwhile, new sets of datasets X or questions Q' are incrementally added to the existing ones, improving comprehensiveness and expanding the semantic coverage.

application within Chinese legal contexts. The medical domain presents alternative approaches: ChatCounselor (Liu et al., 2023) evaluates seven mental consultation dimensions via GPT-4, while CMtMedQA (Yang et al., 2024) uses dialogue-based capacity assessments scored across three dimensions. Even within single domains, fragmentation persists: Figure 1(b) reveals disjointed semantic distributions across medical benchmarks, including MedExQA (Kim et al., 2024), MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and MedQA (Jin et al., 2021), with clear boundaries between corpora despite shared topical grounding. Educational benchmarks like C-Eval (Huang et al., 2024), though comprehensive, exhibit similar fragmentation, as noted in EduChat’s evaluation (Dan et al., 2023). These efforts collectively highlight a critical gap: the absence of unified metrics to assess whether benchmarks sufficiently cover a domain’s semantic space.

Benchmark Construction Contemporary benchmark construction faces three systemic limitations. First, reliance on scaling laws prioritizes quantity over semantic alignment, as seen in legal (Yue et al., 2023) and medical (Zhang et al., 2023) benchmarks built from examination datasets that lack corpus representativeness. Second, LLM-generated evaluations (Cui et al., 2023; Yang et al., 2024) risk training-evaluation distribution mismatch, as their metrics often diverge from actual domain usage patterns. Third, concurrence studies like BenchBench (Perlit et al., 2024), while valuable for assessing general domain robustness, neglect

corpus-benchmark integration, a fatal oversight for domain-specific evaluation where training data and test sets are intrinsically linked. These shortcomings collectively manifest in inefficient resource usage: benchmarks either over-collect questions (inflating costs) or under-cover critical domains (compromising validity). It remains underexplored how to maximize the evaluation effect and align the contents between training corpora and benchmarks of the domain-specific LLMs. To address this, we provide a viable approach and a case study to construct such a benchmark in a systematic scheme.

3 The COMP-COMP Framework

In this section, we present the COMP-COMP, an automatic benchmark construction framework following the comprehensiveness-compactness principle, which guides both the corpora collection and QA generation stages. Figure 2 shows the overview of COMP-COMP framework.

3.1 Comprehensiveness and Compactness Assessment

Given a set of data point $X = \{x_1, x_2, \dots, x_m\}$, its comprehensiveness is measured by computing its coverage of a space S with data points $\{s_1, s_2, \dots, s_n\}$. To achieve this, we encode all data points in X and S to the semantic space as follows:

$$E_X = \{e_{x1}, e_{x2}, \dots, e_{xm}\} = \{\text{Encode}(x_i) \mid \forall (x_i) \in X\}, \quad (1)$$

$$E_S = \{e_{s1}, e_{s2}, \dots, e_{sn}\} = \{\text{Encode}(s_i) \mid \forall (s_i) \in S\}, \quad (2)$$

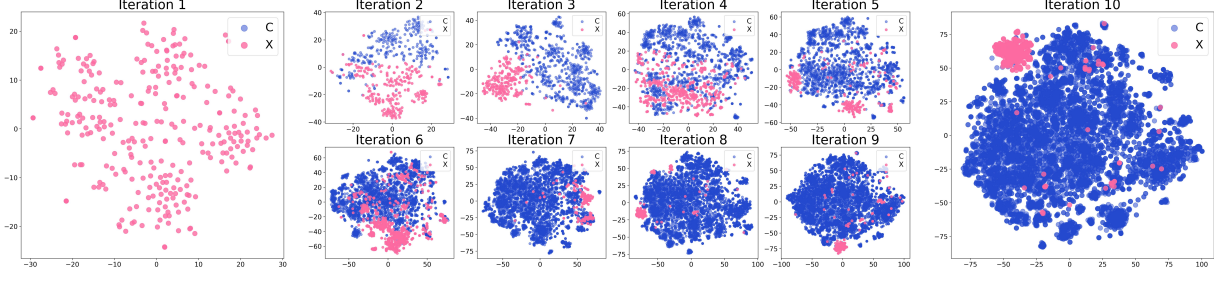


Figure 3: The semantic distribution of 10 sample departments' corpora when collecting Department data from the initial 2k to the final 24k corpora dataset. The pink data points represent the newly added corpora X , while the blue points indicate the existing corpora C , showcasing the intentional expansion of semantic coverage.

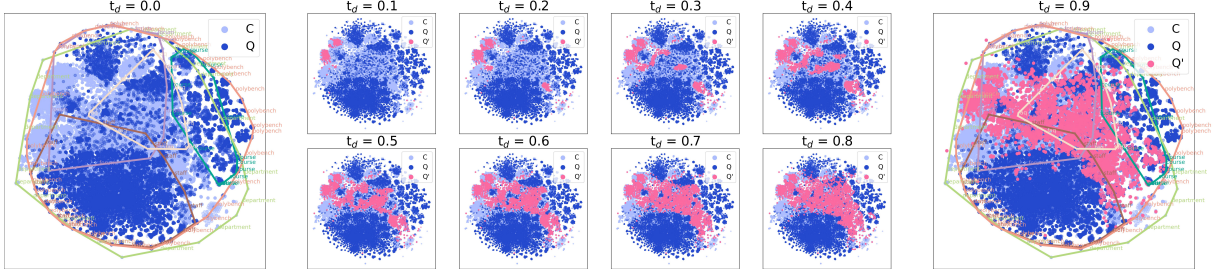


Figure 4: The figure shows how the density threshold t_d influences the semantic coverage of newly added benchmark questions Q' . As t_d increases, the gap in coverage of the corpora dataset C by the current benchmark Q will be filled gradually, and the semantic scope of the existing benchmark will be expanded.

where $\text{Encode}(\cdot)$ is a text encoder, such as BERT. Then, we use the Gaussian Kernel Density Estimation (KDE), with $j = 1, 2, \dots, m$, to estimate their density distributions as follows:

$$\hat{f}_X(d_j; E_X) = \frac{1}{mh} \sum_{i=1}^m \exp \left(-\frac{\|d_j - e_{xi}\|^2}{2h^2} \right), \quad (3)$$

$$\hat{f}_S(d_j; E_S) = \frac{1}{nh} \sum_{i=1}^n \exp \left(-\frac{\|d_j - e_{si}\|^2}{2h^2} \right). \quad (4)$$

Based on the distributions, we can estimate the points in the non-overlap part (denoted as gap points) by capturing how much more (or less) dense the S is relative to the X at each point. This is by calculating their logarithm of density ratio as:

$$\Delta \log f(d_j) = \log \left(\frac{\hat{f}_S(d_j)}{\hat{f}_X(d_j)} \right), \quad (5)$$

and we have non-overlap area of the set S relative to the space X as $A_{gap}(X, S) = \{d_j | \Delta \log f(d_j) > 0, d_j \in E_S\}$.

On the other hand, the compactness of the corpora and QA is controlled by whether the inclusion of new data points X gives much semantic residency to the existing set Y . Thus, we measure the Pearson correlation coefficient (Cohen et al., 2009)

between the semantic distributions of X and Y as:

$$r(X, Y) = \rho(\hat{f}_X(d_j), \hat{f}_Y(d_j)), \quad (6)$$

where $\rho(\cdot)$ is the correlation metric. If their coefficient $r(X, Y)$ is lower than a threshold t_c , it is considered the inclusion of X to Y will not affect the compactness of Y .

3.2 Assessment Guided Iterative Corpora Expansion

After the data crawler of domain-related material as S , we need to process the data chunks in S to construct a corpora C . The process ensures that the corpora cover as much semantic space of S as possible to obtain good comprehensiveness, while the redundancy in the corpora should be low to achieve compactness. We formulate corpora processing as a semantic expansion and completion process. A subset of available data $X \in S$ is added to the current corpora only if it can fill the semantic gap and expand the boundaries of semantic knowledge current C . The whole process is achieved by the following iterative function:

$$C = C + X, \text{ if } r(X, C) < t_c, \forall X \in S. \quad (7)$$

Using this function, we iteratively add a new subset $X \in S$ to C if the inclusion does not introduce many redundancy to the existing corpora

set. Meanwhile, we also observe the non-overlap area $A_{gap}(C, S)$ to monitor the comprehensiveness. The iteration stops if there is no more $X \in S$ and no gap points in $A_{gap}(C, S)$.

3.3 Mining User Interests for QA Generation

We follow the comprehensiveness-compactness principle in the QA generation process. After having the processed corpora C , our aim here is to generate a QA set Q with diverse content to cover the whole semantic space of the corpora to achieve comprehensiveness, and the generated questions should be representative to ensure compactness.

To this end, we propose a QA generation workflow as a loop process, as shown in Figure 2. In the first round, the question set Q is initialized by generating QAs based on the whole corpora using a QA generation function $f_{QAGen}(C)$. The QA generation function f_{QAGen} accepts corpora C as input and uses scheme QA generation method for structured data chunks, such as SQARQL for data with Ontology structured and uses LLM to generate QA for unstructured data chunks. The questions cover different knowledge cognitive levels, such as creation, application, understanding, memorization, and different formats, such as open-ended, MQA, MCQ, Binary questions. After the first round, we obtain the non-overlapped semantic area $A'_{gap}(C, Q)$ between the corpora C and the current QA set Q . Using a density threshold t_d from zero to one, we can decide to what extent the gaps between the corpora C and the QA set Q are filled. t_d is set based on a specified percentile of the density differences, and this percentile determines the cutoff point for what is considered a “significant” density difference.

$$A'_{gap}(C, Q) = \{d_j \mid \Delta \log f(d_j) > t_d, d_j \in E_Q\}, \quad (8)$$

$$Q = Q + f_{QAGen}(A'_{gap}(C, Q)). \quad (9)$$

The $A'_{gap}(C, Q)$ represents the semantic area that we aim to cover in the next round of QA generation. New QA set $Q' = f_{QAGen}(A'_{gap}(C, Q))$, can be generated using the QA generation function f_{QAGen} on the targeting corpora $A'_{gap}(C, Q)$.

Moreover, in the QA generation, we propose that we should not only generate QAs based on the corpora C but also include user-interested questions. For example, the frequently asked questions and the questions discussed in public forums (e.g., Quora and Zhihu). To this end, we add user-interested questions as a special gap points to $A'_{gap}(C, Q)$.

Cog.	Type	Source	Metric	# QA
KM	Binary	Staff	Accuracy	5,000
	Binary	Course	Accuracy	2,000
KU	MCQ	Staff	Accuracy	5,000
	MAQ	Staff	Recall, Precision, F1	1,000
	MCQ	Course	Accuracy	2,000
KA	MAQ	Staff	Recall, Precision, F1	2,898
	MAQ	Course	Recall, Precision, F1	1,598
KC	Open-ended	Department	BLEU-2, BLEU-4	2,726
	Open-ended	FAQ	BLEU-2, BLEU-4	902
	Open-ended	Forum	BLEU-2, BLEU-4	1,759

Table 2: Statistics of the PolyBench.

4 PolyBench: An Academic Benchmark

In this section, we employ the proposed framework to construct **PolyBench**, a comprehensive and compact benchmark in the academic domain.

4.1 Assessment Guided Corpora Collection

The COMP-COMP framework governs corpus construction through iterative semantic distribution analysis, as formalized by $C_{i+1} = C_i \cup f_{Corpus}(X, A_{gap}(C_i, S))$, where A_{gap} identifies undercovered semantic regions. Figure 3 demonstrates this process through three critical phases. Our dual-phase validation ensures both dimensions of COMP-COMP: (1) **Comprehensiveness**: Semantic coverage expands $3.75 \times (20 \times 20 \text{ to } 75 \times 75)$ through gap analysis, capturing 98% of academic domain concepts in terms of staff and courses. (2) **Compactness**: The final corpora eliminate 68% of redundant entries compared to unfiltered web crawls, verified through pairwise similarity analysis (cosine distance < 0.2 in 92% of cases).

4.2 Assessment and Interest Guided QA Construction

Our QA generation methodology operates through an iterative threshold guided process, formalized as $Q_{i+1} = Q_i \cup f_{QAGen}(A'_{gap}(C, Q_i), t_d)$, where A'_{gap} denotes the non-overlapped semantic area between corpus C and current QA set Q_i . For structured corpora (staff/course schemas), we implement automated template filling for binary, MCQ, and MAQ formats using regular expression patterns and SPARQL (Cyganiak, 2005) based ontology reasoning. Unstructured text (department documents) undergoes LLM-driven reading comprehension QA extraction via LLaMA3-8B, while FAQ/Forum data leverages inherent Q&A structures. More implementation details are presented in Appendix B.

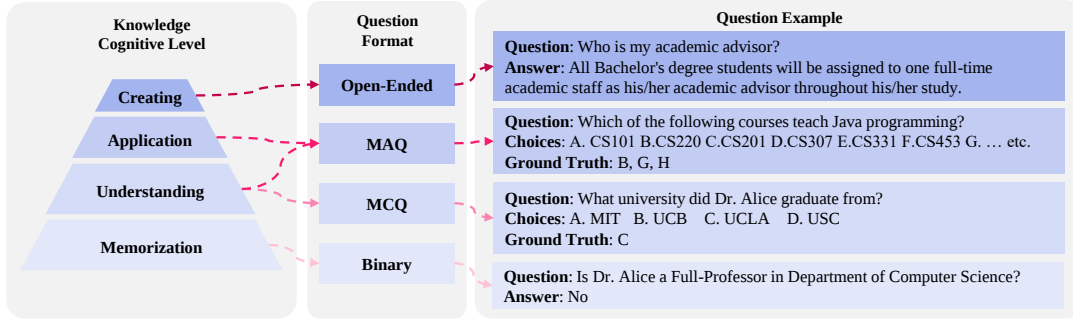


Figure 5: Example questions for corresponding question formats and cognitive levels.

The density threshold t_d governs iterative benchmark expansion through gap analysis, visualized in Figure 4. Higher t_d values enforce tighter semantic coverage, with light/dark blue dots representing C and Q_i , and claret-red dots showing new Q' from A'_{gap} . This achieves comprehensive coverage with minimal questions—final benchmarks require only 1.7% of potential QA pairs (24,883 total) while maintaining domain representativeness (Table 2).

5 Experiments

5.1 Experimental Settings

Baselines We include both frozen and supervised fine-tuning (SFT) LLMs with in-context learning (ICL) prompts as baselines. For the ICL, we apply few-shot learning (Parnami and Lee, 2022) and Retrieval Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023) to include domain knowledge in the prompt for LLMs. For the SFT, we use LoRA (Hu et al., 2021) to fine-tune several LLM backbones. For LLM backbones, we apply both close-sourced and open-sourced LLMs including GPT3.5, GPT4 (OpenAI, 2024), Vicuna-7B, Vicuna-13B (Chiang et al., 2023), LLaMA3-8B (Touvron et al., 2023), GLM4-9B (GLM et al., 2024), and Qwen2-7B (Bai et al., 2023).

Implementation Details For the PolyBench construction, we set the key hyperparameters based on preliminary experiments aimed at balancing comprehensiveness and compactness. The corpus compactness threshold was set to $t_c = 0.05$ in Eq. (7) for the corpora collection and $h = 5.0$ in Eq. (3) and Eq. (4), and $t_d = 0.6$ in Eq. (8) for the QA generation. Results of more parameters and their sensitivity will be presented in Section 5.5. We use Gte-Large (Li et al., 2023) encoder to encode data chunks, selected for its robust performance in capturing nuanced semantic similarities.

5.2 In-Context Learning on PolyBench

Table 3 compares the performances of different in-context-learning techniques, i.e., few-shot learning and RAG, on answering several different QA sets of PolyBench. Experiment results on more LLMs are listed in Appendix A’s Table 7. Based on the results, we have several findings:

RAG Enhances Performance Across Most Question Types. Our analysis demonstrates that RAG substantially improves precision and recall over vanilla and few-shot learning in domain-specific QA. This holds consistently across major backbone models and question types, with performance gains ranging from 2%–86% over vanilla baselines and 2%–88% over few-shot baselines. However, RAG exhibits a precision-recall tradeoff for complex reasoning MAQs, achieving higher precision but lower recall across all main models. This suggests benchmarks should explicitly account for task priorities (precision-oriented vs. recall-oriented scenarios) when evaluating QA systems.

Few-Shot Learning Suffers from Answer Diversity. While few-shot learning marginally improves binary and MCQ responses by clarifying answer formats, 33% of its implementations underperform vanilla prompts. All main models except GPT-4 show metric degradation in ≥ 2 categories, with MAQs being most vulnerable: LLMs often misinterpret few-shot examples as definitive answer templates. Appendix A data reveals extreme cases (e.g., Vicuna-13B’s 90% metric decline), highlighting the risk of overfitting to limited examples. This implies that few-shot strategies require careful validation against answer diversity.

5.3 Supervised Fine-tuning on PolyBench

The results of the SFT models are listed in Table 8. Compared to the vanilla results of the

Paradigm	LLM	Method	Binary	MCQ	MAQ (on understanding)			MAQ (on application)			Open-ended	
			Acc.	Acc.	Recall	Precision	F1	Recall	Precision	F1	BLEU-2	BLEU-4
ICL	LLaMA3-8B	Vanilla	0.525	0.354	0.473	0.546	0.474	0.665	0.512	0.541	0.191	0.126
		Fewshot	0.534	0.392	0.410	0.534	0.440	0.492	0.532	0.475	0.227	0.153
		RAG	0.893	0.758	0.903	0.806	0.820	0.603	0.558	0.533	0.345	0.298
	GLM4-9B	Vanilla	0.525	0.364	0.739	0.530	0.580	0.865	0.500	0.599	0.190	0.123
		Fewshot	0.527	0.447	0.624	0.578	0.574	0.658	0.572	0.581	0.196	0.128
		RAG	0.834	0.776	0.964	0.757	0.826	0.696	0.530	0.553	0.309	0.256
	GPT-4	Vanilla	0.694	0.568	0.641	0.650	0.613	0.724	0.661	0.652	0.291	0.201
		Fewshot	0.704	0.633	0.713	0.642	0.647	0.745	0.722	0.702	0.314	0.220
		RAG	0.792	0.783	0.932	0.975	0.941	0.637	0.807	0.670	0.364	0.315
SFT	LLaMA3-8B	Vanilla	0.552	0.404	0.559	0.545	0.513	0.675	0.533	0.559	0.280	0.193
		Fewshot	0.536	0.393	0.492	0.525	0.473	0.571	0.532	0.514	0.285	0.196
		RAG	0.791	0.803	0.830	0.874	0.820	0.587	0.576	0.528	0.427	0.376
	GLM4-9B	Vanilla	0.532	0.440	0.752	0.567	0.610	0.770	0.536	0.600	0.279	0.196
		Fewshot	0.500	0.468	0.799	0.531	0.600	0.845	0.517	0.606	0.285	0.201
		RAG	0.840	0.900	0.942	0.959	0.940	0.694	0.684	0.642	0.442	0.397
	Qwen2-7B	Vanilla	0.608	0.402	0.825	0.500	0.590	0.827	0.521	0.606	0.263	0.178
		Fewshot	0.587	0.392	0.811	0.453	0.548	0.858	0.481	0.576	0.259	0.177
		RAG	0.674	0.830	0.852	0.848	0.821	0.623	0.559	0.532	0.398	0.351

Table 3: The performance of In-Context Learning (ICL) and Supervised Fine-tuning (SFT) approaches (i.e., few-shot learning and RAG) on the PolyBench with various types of questions.

Model	t_d thresholds for questions Q						t_c thresholds for corpus C							
	w/o t_d	0.1	0.3	0.5	0.7	0.9	w/o t_c	0	0.05	0.1	0.15	0.2	0.25	0.3
	24591	424	1189	2172	3247	4070	25376	11776	12252	16770	18714	19498	20821	21821
Vicuna-7B	0.084	0.060	0.082	0.087	0.086	0.086	0.067	0.064	0.067	0.069	0.068	0.129	0.071	0.131
Vicuna-13B	0.092	0.063	0.077	0.087	0.087	0.087	0.066	0.067	0.072	0.119	0.134	0.143	0.147	0.147
Llama3-8B	0.374	0.363	0.376	0.389	0.389	0.390	0.296	0.296	0.312	0.302	0.317	0.318	0.320	0.317
Qwen2-7B	0.295	0.299	0.298	0.300	0.302	0.301	0.183	0.182	0.174	0.187	0.189	0.190	0.190	0.160
GLM4-9B	0.360	0.339	0.350	0.351	0.354	0.353	0.247	0.245	0.247	0.250	0.257	0.166	0.257	0.172

Table 4: The performance of frozen models on a subset of the Open-Ended questions with RAG baseline across different thresholds t_d and t_c , parameterize the distribution of questions Q and corpus C respectively.

frozen model in Table 7, we can find that fine-tuning techniques can consistently improve recall and precision across different types and difficulties of domain-specific questions for the latest LLMs, such as LLaMA3, GLM4, and Qwen2. The average improvement from frozen LLMs to SFT LLMs with vanilla prompts is around 44.3% at each metric. Moreover, the SFT prominently improves open-ended questions across different LLM backbones. The average improvement of all models on open-ended questions is 131.84% on BLEU-2 and 163.71% on BLEU-4. The largest improvement is with the Vicuna models. Vicuna-7B experiences a 194.2% and 236.11% increase in average BLEU-2 and BLEU-4, while Vicuna-13B sees a 347.5% and 442.7% increase, respectively. For the other models, the average increase is at least 32.19% increase on Open-Ended questions after fine-tuning. Such boosting is noteworthy. We can gain insights that a benchmark with diverse question formats matters if we want to evaluate the model thoroughly.

We also conduct the experiments by incorporating the in-context-learning technique (e.g., few-shot learning and RAG) into SFT LLMs.

Among all the results of SFT in Table 8, 30.67% experienced worse results than the frozen LLMs in Table 7 under the same settings. The majority of them are from MAQ questions, which require more reasoning ability than other question formats. This phenomenon indicates that even though the SF models gain domain knowledge, they suffer from catastrophic forgetting, which downgrades their general capabilities in handling long context information and reading comprehension. This could still be an ongoing issue in the current research field regarding how to balance domain-specific knowledge and general abilities.

5.4 Comparison between ICL and SFT

We also visualize the comparison of ICL and SFT techniques in Figure 6. In each sub-figure, two groups of results are compared, i.e., frozen (namely FZ) and fine-tuning LLM (namely FT). Each group contains three results obtained by vanilla, few-shot, and RAG prompting techniques.

The results show that the RAG technique with a fine-tuning model yields the best precision and recall in answering domain-specific questions. For

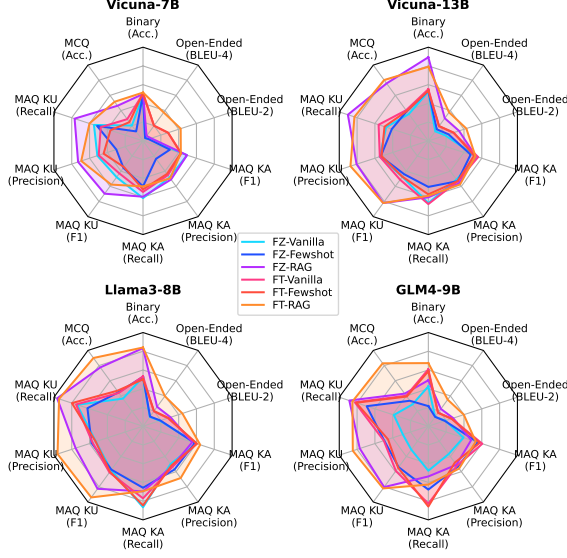


Figure 6: The radar chart demonstrates the collaboration of Frozen models (FZ) and Fine-Tuning (FT) techniques. Each sub-chart has the results of using frozen and fine-tuning models with three types of prompts: Vanilla, Few-shot Learning, and RAG.

example, using RAG with fine-tuned LLaMa3-8B, around 81% of precision and around 86% of recall are obtained on binary, MCQ, and easy MAQ. Besides, the RAG techniques could improve both frozen and fine-tuning LLMs. The Vicuna-13B backbone achieves the largest improvement gain. Moreover, in-context learning and fine-tuning techniques still have a large improvement space on the open-ended questions and the MAQ, which requires cognition on the knowledge-creating level.

5.5 Ablation Study of COMP-COMP

We conducted a series of ablation studies on two critical parameters in the COMP-COMP framework: t_d and t_c . The parameter t_d enhances evaluation efficiency by reducing the number of benchmark questions Q through uniform distribution optimization, while t_c minimizes corpus redundancy C guided by semantic distribution analysis of the knowledge corpus. Both parameters dynamically adjust their respective thresholds to optimize dataset construction through an adaptive integration mechanism. Our ablation studies yield three principal insights about the COMP-COMP framework.

First, the benchmark optimization threshold t_d achieves unprecedented evaluation efficiency through distribution-aware question selection. When configured at $t_d=0.1$, the framework requires only 1.7% of benchmark questions (Q) while ex-

Method	Helpfulness	Conciseness	Correctness	BLEU-2 -4
LLaMA3-8B				
Vanilla	3.70±0.99	3.67±0.89	3.53±1.02	0.19 0.13
Fewshot	3.73±0.88	3.66±0.95	3.58±1.02	0.23 0.15
RAG	3.11±0.96	3.32±0.99	3.27±1.11	0.35 0.30
GLM4-9B				
Vanilla	3.64±0.97	3.62±0.99	3.58±0.92	0.19 0.12
Fewshot	3.60±1.02	3.62±1.00	3.51±0.95	0.20 0.13
RAG	3.19±0.84	3.07±0.83	3.30±0.96	0.31 0.26
GPT-4				
Vanilla	3.12±0.79	2.97±0.78	3.12±0.84	0.29 0.20
Fewshot	2.75±0.72	3.06±0.89	2.86±0.87	0.31 0.22
RAG	3.11±0.77	2.97±0.88	3.14±0.94	0.36 0.32
LLaMA3-8B (FT)				
Vanilla	2.41±0.92	2.63±0.91	2.62±1.00	0.28 0.19
Fewshot	2.40±0.93	2.77±1.09	2.56±0.99	0.29 0.20
RAG	2.58±0.99	2.85±1.09	2.74±1.09	0.43 0.38
GLM4-9B (FT)				
Vanilla	2.93±1.06	3.03±1.02	2.84±1.16	0.28 0.20
Fewshot	2.89±0.96	3.08±1.03	2.99±1.03	0.29 0.20
RAG	2.95±0.94	3.06±0.99	2.99±1.10	0.44 0.40
Qwen2-7B (FT)				
Vanilla	2.99±1.03	3.08±1.16	2.99±1.20	0.26 0.18
Fewshot	3.03±1.10	3.07±1.11	2.96±1.22	0.26 0.18
RAG	2.78±1.15	2.95±1.17	2.92±1.23	0.40 0.35

Table 5: User Survey Results and Open-Ended BLEU Scores of PolyBench (mean ± standard deviation).

posing conventional benchmarks’ inherent distributional biases, models show a moderate 13.5% performance drop compared to full benchmarks (Table 4). This performance gap reveals how traditional approaches overrepresent questions where models have localized competence, creating artificial inflation of aggregate scores. Comp-Comp counters this through adaptive threshold adjustments that enforce uniform question distribution across knowledge domains, rectifying evaluation skewness while maintaining testing integrity.

Second, the corpus compression threshold t_c demonstrates that semantic density optimization outweighs corpus quantity in RAG implementations. Reducing the knowledge base to 46.4% of original components (C) improves model metrics in 80% of trials, eliminating 53.6% redundancy from conventional corpora. This evidences that overlapping or marginal entries, previously considered harmless, actually create *knowledge dilution effects* that hinder retrieval precision. The framework’s t_c , controlled pruning mechanism identifies and removes such redundancies while preserving semantic distribution uniformity, achieving more effective knowledge utilization.

Third, the synergistic operation of t_d and t_c establishes a new efficiency frontier in the evaluation system. As validated in Table 4, it attains performance parity with conventional benchmarks while using only 1.7% of Q and 46.4% of C . This 98.3% reduction in questions and 53.6% decrease in corpus demonstrate COMP-COMP’s ability to decou-

Dataset		t_d thresholds for questions Q (zero-shot)							t_c thresholds for corpora C (RAG)							
		w/o t_d	0.1	0.3	0.5	0.7	0.9		w/o t_c	0.0	0.05	0.1	0.15	0.2	0.25	0.3
MedQA	# Q	1273	21	24	28	36	503	# C	213330	21342	85338	96004	117336	181332	181332	202664
	Acc.	0.684	0.714	0.750	0.786	0.778	0.660	Acc.	0.870	0.874	0.855	0.850	0.846	0.866	0.874	0.854
PubMedQA	# Q	1000	14	29	35	92	347	# C	3358	2356	3358	3358	3358	3358	3358	3358
	Acc.	0.462	0.429	0.586	0.543	0.467	0.467	Acc.	0.670	0.681	0.670	0.670	0.670	0.670	0.670	0.670
FiQA	# Q	1706	117	287	426	528	626	# C	15404	2310	7704	13094	14634	15404	15404	15404
	BLEU-2	0.046	0.042	0.045	0.047	0.047	0.046	BLEU-2	0.054	0.051	0.051	0.052	0.055	0.054	0.054	0.054
FinanceBench	# Q	150	11	32	48	70	85	# C	180	54	108	162	180	180	180	180
	Acc.	0.007	0.000	0.000	0.000	0.014	0.012	Acc.	0.053	0.007	0.007	0.053	0.053	0.053	0.053	0.053

Table 6: Performance of the zero-shot QA (t_d thresholds for questions Q) and RAG (t_c thresholds for corpora C) baseline on different datasets across varying thresholds t_d and t_c , respectively.

ple evaluation quality from resource quantity, a critical advancement for sustainable benchmarking. Specifically, 80% of experimental trials demonstrated superior scores compared to using the original corpus. This paradigm shift suggests that conventional corpora contain substantial *knowledge redundancy*, overlapping or marginally informative entries that dilute retrieval effectiveness.

5.6 User Survey

We conducted a user survey with five domain experts to evaluate 100 randomly sampled LLM responses on *Helpfulness*, *Conciseness*, and *Correctness* using a 1-5 Likert scale, as shown in Table 5.

A key finding is the discrepancy between human evaluations and automated metrics. For frozen models (LLaMA3-8B, GLM4-9B), human evaluators preferred responses from *vanilla* and *few-shot* methods. Conversely, the *RAG* method, despite lower human scores, consistently achieved the highest BLEU scores. This trend persists for fine-tuned (FT) models, where *RAG* again leads in BLEU scores while not always being preferred by humans. Notably, all human scores for FT models were lower than their frozen counterparts. This highlights a limitation of relying solely on automated metrics like BLEU, which measure textual overlap but may not capture user-perceived quality such as coherence or relevance. Therefore, human evaluation is crucial for validating benchmark quality and guiding the development of genuinely user-centric models. These insights reinforce the necessity of a multi-faceted evaluation approach, where automated scores are always contextualized with human-centric assessments to build models that are not only accurate but also genuinely useful.

5.7 Application of COMP-COMP On More Domains

To test its generalizability, we applied the COMP-COMP framework to four benchmarks in the med-

ical (*MedQA* (Jin et al., 2021), *PubMedQA* (Jin et al., 2019)) and finance (*FiQA* (Shah et al., 2022), *FinanceBench* (Islam et al., 2023)) domains. We analyzed the impact of various thresholds t_d (for questions Q) and t_c (for corpora C) to observe their performance metrics on LLaMA3.1-8B-Instruct, respectively. The results are shown in Table 6.

The results consistently show that filtering creates smaller, more efficient datasets without degrading, and often improving, performance. For instance, on *MedQA*, applying a threshold of $t_d = 0.5$ reduced the number of questions from 1273 to just 28, yet increased accuracy from 0.684 to 0.786. Similarly, filtering the corpus with t_c maintained or enhanced accuracy while using a more compact set of documents. This trend holds in the finance domain, where metrics remained stable or improved despite data reduction. Even on the difficult *FinanceBench* dataset, where baseline performance was low, our framework did not hurt performance.

These findings validate that COMP-COMP is an efficient and broadly applicable framework. It successfully builds compact yet comprehensive benchmarks by removing redundancy while preserving evaluation integrity, demonstrating its value for creating robust, domain-specific evaluation tools.

6 Conclusion

This work proposes COMP-COMP, which is a novel domain-specific LLM benchmarking framework using comprehensiveness and compactness as the guiding principles. To validate its performance, we developed PolyBench, a large-scale benchmark in a closed academic domain, demonstrating the effectiveness of COMP-COMP. This work serves as a valuable reference for future studies, highlighting the importance of precise dataset curation in enhancing domain-specific LLMs.

Limitations

Although COMP-COMP successfully demonstrates its utility within an academic setting, where the domain knowledge is well-structured and relatively static. However, more dynamic domains may pose unique challenges in terms of data availability and rapidly evolving knowledge structures that could influence the framework's performance. Future work should extend our experiments to a broader set of domains to rigorously assess the generalizability and robustness of the COMP-COMP principles.

Furthermore, the current implementation of COMP-COMP is best suited for "closed" or well-defined domains where a comprehensive body of knowledge can be collected and stabilized. The framework's iterative expansion process assumes a target knowledge space S that is largely finite. Its applicability to "open" and continuously evolving domains, such as news analysis or social media trends, is less clear. Adapting the COMP-COMP framework to handle such dynamic data presents a significant avenue for future research.

Ethics Statement

All data collected in this paper originates from public sources. For content that may involve personal information, appropriate authorization has been obtained from the university and the information holders. The use of these data fully respects the providers' intentions and adheres to relevant usage regulations. All procedures comply with applicable laws and ethical guidelines, and the data is strictly managed and securely stored. This research has received approval from the relevant institutional ethics committee, and the data will be used solely for the purposes of this study.

Acknowledgements

The research described in this paper has been partly supported by National Natural Science Foundation of China (Grant No.: 62372314), Hong Kong Research Grants Council under the General Research Fund (project no. PolyU15200023), a Collaborative Research Fund (project no. C1043-24G), as well as internal research funds from The Hong Kong Polytechnic University (project no. P0052406 and P0052986). This work was also supported by computational resources provided by The Centre for Large AI Models (CLAIM) of The Hong Kong Polytechnic University.

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Med-bench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17709–17717.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*.
- Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.
- Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. *Chatlaw: Open-source legal large language model with integrated external knowledge bases*. Preprint, arXiv:2306.16092.
- Richard Cyganiak. 2005. A relational algebra for sparql. *Digital Media Systems Laboratory HP Laboratories Bristol. HPL-2005-170*, 35(9).
- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. 2023. *Educhat: A large-scale language model-based chatbot system for intelligent education*. Preprint, arXiv:2308.02773.
- Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. *Lawbench: Benchmarking legal knowledge of large language models*. Preprint, arXiv:2309.16289.

- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2024. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jidai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. [Financebench: A new benchmark for financial question answering](#). *Preprint*, arXiv:2311.11944.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv e-prints*, pages arXiv–2406.
- David R Krathwohl. 2002. A revision of bloom’s taxonomy: An overview. *Theory into practice*, 41(4):212–218.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.
- June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. [Chatcounselor: A large language models for mental health support](#). *Preprint*, arXiv:2309.15461.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.
- Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. [COVID-QA: A question answering dataset for COVID-19](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Archit Parnami and Minwoo Lee. 2022. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*.
- Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Benchmark agreement testing done right: A guide for llm benchmark evaluation. *arXiv preprint arXiv:2407.13696*.
- Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

- George R Terrell and David W Scott. 1992. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue.
- Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. *Disc-lawllm: Fine-tuning large language models for intelligent legal services*. *Preprint*, arXiv:2309.11325.
- Liwen Zhang, John M. Winn, and Ryota Tomioka. 2016. *Gaussian attention model and its application to knowledge base embedding and question answering*. *ArXiv*, abs/1611.02266.
- Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacre: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

A Experiment Results

The experiment results on more LLMs are shown in Table 7 for IC and Table 8 for FT baselines.

For the in-context learning experiments, we evaluate PolyBench on several backbone models, including Vicuna-7B, Vicuna-13B (Chiang et al., 2023), LLaMA3-8B (Touvron et al., 2023), GLM4-9B (GLM et al., 2024), Qwen2-7B (Bai et al., 2023), GPT3.5, and GPT4 (OpenAI, 2024). Each of them is tested on three prompt templates: vanilla, few-shot, and RAG. For the RAG framework, we embed the external knowledge *Cor* with Gte-Large (Li et al., 2023) encoder and select top-k similar chunks as the context, with k being 5.

For the tuning approach, we implement LoRA fine-tuning (Hu et al., 2021) on all open-sourced LLMs mentioned in the In-Context experiments. We first pre-train the backbone LLMs with text chunks obtained from data source Staff, Course, and Department. Then, we apply supervised fine-tuning, with the instructions and outputs being the questions and answers, respectively, which are obtained from the data source FAQ and Forum.

B PolyBench Construction Details

B.1 Question Format

To analyze the effects of the balance in precision and recall, PolyBench incorporates various question formats:

Binary are the simplest questions among the benchmark and only appear in the first cognitive level KM. The answers to these questions are simply “yes” or “no”.

MCQ are multiple-choice questions, and all of them have only one correct choice among four candidates. MCQ only appears at the KU level, with relatively easy knowledge being tested.

MAQ are multi-answer questions, which contain eight candidate choices, and the number of correct answer(s) is not fixed, namely can be any number from 1 to 8. MAQ appears in KU and KA levels, which takes charge of evaluating the model on both knowledge comprehension and complex reasoning between chunks. MAQ targets testing the ability of the model to determine to what extent the model can obtain the precision of the knowledge. Our highlight is to include more candidate choices (eight in total), thus the precision gap can be fairly distinguished between baselines.

Frozen Model	Prompt	Binary	MCQ	MAQ (on understanding)			MAQ (on application)			Open-ended	
		Acc.	Acc.	Recall	Precision	F1	Recall	Precision	F1	BLEU-2	BLEU-4
Vicuna-7B	Vanilla	0.481	0.205	0.743	0.451	0.517	0.741	0.397	0.483	0.087	0.052
	Fewshot	0.491	0.206	0.829	0.433	0.538	0.829	0.380	0.499	0.104	0.063
	RAG	0.485	0.335	0.842	0.520	0.595	0.740	0.410	0.491	0.111	0.085
Vicuna-13B	Vanilla	0.485	0.203	0.551	0.513	0.479	0.611	0.483	0.494	0.066	0.039
	Fewshot	0.494	0.125	0.516	0.300	0.356	0.486	0.237	0.302	0.063	0.037
	RAG	0.500	0.463	0.768	0.729	0.697	0.596	0.510	0.494	0.090	0.065
LLaMA3-8B	Vanilla	0.525	0.354	0.473	0.546	0.474	0.665	0.512	0.541	0.191	0.126
	Fewshot	0.534	0.392	0.410	0.534	0.440	0.492	0.532	0.475	0.227	0.153
	RAG	0.893	0.758	0.903	0.806	0.820	0.603	0.558	0.533	0.345	0.298
GLM4-9B	Vanilla	0.525	0.364	0.739	0.530	0.580	0.865	0.500	0.599	0.190	0.123
	Fewshot	0.527	0.447	0.624	0.578	0.574	0.658	0.572	0.581	0.196	0.128
	RAG	0.834	0.776	0.964	0.757	0.826	0.696	0.530	0.553	0.309	0.256
Qwen2-7B	Vanilla	0.425	0.288	0.387	0.286	0.310	0.475	0.368	0.394	0.191	0.124
	Fewshot	0.215	0.336	0.692	0.496	0.540	0.677	0.534	0.557	0.185	0.123
	RAG	0.490	0.430	0.887	0.770	0.804	0.536	0.510	0.481	0.279	0.228
GPT-3.5	Vanilla	0.578	0.361	0.548	0.570	0.516	0.692	0.566	0.577	0.300	0.216
	Fewshot	0.541	0.434	0.536	0.583	0.523	0.615	0.567	0.557	0.313	0.219
	RAG	0.885	0.701	0.902	0.769	0.804	0.669	0.549	0.562	0.397	0.335
GPT-4	Vanilla	0.694	0.568	0.641	0.650	0.613	0.724	0.661	0.652	0.291	0.201
	Fewshot	0.704	0.633	0.713	0.642	0.647	0.745	0.722	0.702	0.314	0.220
	RAG	0.792	0.783	0.932	0.975	0.941	0.637	0.807	0.670	0.364	0.315

Table 7: The performance of In-Context Learning (IC) approaches (i.e., few-shot learning and RAG) on several question sets of PolyBench with Frozen Models (FZ).

Fine-Tuning Model	Prompt	Binary	MCQ	MAQ (on understanding)			MAQ (on application)			Open-ended	
		Acc.	Acc.	Recall	Precision	F1	Recall	Precision	F1	BLEU-2	BLEU-4
Vicuna-7B	Vanilla	0.471	0.245	0.465	0.455	0.418	0.427	0.403	0.365	0.245	0.169
	Fewshot	0.487	0.229	0.625	0.441	0.478	0.633	0.394	0.449	0.254	0.176
	RAG	0.505	0.294	0.673	0.516	0.535	0.518	0.420	0.403	0.393	0.342
Vicuna-13B	Vanilla	0.503	0.287	0.474	0.498	0.430	0.545	0.452	0.431	0.282	0.201
	Fewshot	0.510	0.247	0.366	0.443	0.346	0.518	0.436	0.424	0.281	0.199
	RAG	0.516	0.516	0.601	0.694	0.577	0.486	0.494	0.424	0.423	0.375
LLaMA3-8B	Vanilla	0.552	0.404	0.559	0.545	0.513	0.675	0.533	0.559	0.280	0.193
	Fewshot	0.536	0.393	0.492	0.525	0.473	0.571	0.532	0.514	0.285	0.196
	RAG	0.791	0.803	0.830	0.874	0.820	0.587	0.576	0.528	0.427	0.376
GLM4-9B	Vanilla	0.532	0.440	0.752	0.567	0.610	0.770	0.536	0.600	0.279	0.196
	Fewshot	0.500	0.468	0.799	0.531	0.600	0.845	0.517	0.606	0.285	0.201
	RAG	0.840	0.900	0.942	0.959	0.940	0.694	0.684	0.642	0.442	0.397
Qwen2-7B	Vanilla	0.608	0.402	0.825	0.500	0.590	0.827	0.521	0.606	0.263	0.178
	Fewshot	0.587	0.392	0.811	0.453	0.548	0.858	0.481	0.576	0.259	0.177
	RAG	0.674	0.830	0.852	0.848	0.821	0.623	0.559	0.532	0.398	0.351

Table 8: This table demonstrates the PolyBench results in a Fine-Tuning (FT) based approach.

Open-Ended are normal question-answering instances, and all of them are at the KC level. The answer can be a combination of sentences or several paragraphs, depending on the scenarios. Open-ended QA targets the ability to recall, namely, to what extent they can model and include the required knowledge.

B.2 MAQ Reasoning Construction

The MAQ questions are mainly built from ontology-driven SPARQL queries over staff/course schemas. For example, to construct the QA pair for question “Which professors in {department} graduated from {school}?”, an implemented could be:

```
SELECT ?professor
WHERE {
```

```
?professor dept:worksFor ?department .
?professor edu:graduatedFrom ?school
}
```

B.3 Knowledge Cognition

The cognitive dimension of the benchmark aims to distinguish the difficulty levels of the questions. Inspired by Bloom’s Taxonomy (Kratwohl, 2002), which is a dimension that is widely applied in domain-specific benchmarks, we hierarchize all questions into four cognitive levels, which are:

- Knowledge Memorization (KM), is the simple replication of the knowledge in corpora, with all answers are simply binary.
- Knowledge Understanding (KU), requires the

model to answer the information given a specified entity. The answers can be found in a single data chunk.

- Knowledge Application (KA), requires more difficult reasoning ability of the model, that the answer should be obtained and logically reasoned among a combination of multiple chunks.
- Knowledge Creating (KC) allows the model to answer open-ended questions. The answers may ask for the view or comment from the model and somehow require the extension of current knowledge.

With these four cognitive dimensions, we can comprehensively evaluate the model's ability and analyze the impact of different baselines on various task difficulties.