

SDBench: A Survey-based Domain-specific LLM Benchmarking and Optimization Framework

Cheng Guo¹, Kai Hu¹, Shuxian Liang^{2,3}, Yiyang Jiang¹,
Yi Gao¹, Xian-Sheng Hua^{2,3}, Wei Dong¹

¹Zhejiang University ²Terminus Group ³Tongji University
{guo.cheng,hk_zju,jiangyiyang,gaoyi,dongw}@zju.edu.cn,
shuxianliang0304@gmail.com, xshua@tongji.edu.cn

Abstract

The rapid advancement of large language models (LLMs) in recent years has made it feasible to establish domain-specific LLMs for specialized fields. However, in practical development, acquiring domain-specific knowledge often requires a significant amount of professional expert manpower. Moreover, even when domain-specific data is available, the lack of a unified methodology for benchmark dataset establishment often results in uneven data distribution. This imbalance can lead to an inaccurate assessment of the true model capabilities during the evaluation of domain-specific LLMs. To address these challenges, we introduce **SDBench**, a generic framework for generating evaluation datasets for domain-specific LLMs. This method is also applicable for establishing the LLM instruction datasets. It significantly reduces the reliance on expert manpower while ensuring that the collected data is uniformly distributed. To validate the effectiveness of this framework, we also present the **BridgeBench**, a novel benchmark for bridge engineering knowledge, and the **BridgeGPT**, the first LLM specialized in bridge engineering, which can solve bridge engineering tasks.

1 Introduction

With the widespread adoption and impact of GPT-3 (Brown et al., 2020), Large Language Models (LLMs) have witnessed remarkable advancement in recent years. New models such as Llama3 (Touvron et al., 2023), Qwen2.5 (Yang et al., 2024), GPT-4 (Achiam et al., 2023), Claude3 (Anthropic, 2024) and Deepseek-R1 (Guo et al., 2025) continue to achieve state-of-the-art performance across various benchmarks. However, current LLMs, primarily trained on public knowledge corpus, often fail to meet the requirements of domain experts who work with private offline knowledge. Moreover, deploying domain-specific LLM typically necessitates on-premise model de-

ployment. Hence, training and deploying domain-specific LLM with fewer parameter counts remain a more viable solution for numerous domain experts.

During the early years of domain-specific applications, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) emerged as an efficient approach among researchers to ensure response accuracy. In recent years, with the diversification of domain applications and the maturation of LLMs, retraining domain-specific LLMs has become another viable option. Current domain-specific LLMs have achieved notable success primarily in some domains, such as biology, healthcare, geography, etc. These domain-specific LLMs benefit from dedicated research organizations that maintain well-structured knowledge architectures and systematically organized datasets, with substantial contributions from a large number of domain experts. The majority of these datasets are publicly accessible through the internet resources. Based on well-structured knowledge, recent work such as OceanGPT (Bi et al., 2023) has proposed a feasible workflow for constructing domain-specific LLMs. This involves defining the domain knowledge architecture through expert panels, creating datasets with the collaboration of experts and data engineers, and ultimately generating corresponding instruction datasets (Ocean Instruction Dataset) and benchmark datasets (OceanBench) for training and evaluation.

However, the development of a novel domain-specific large language model presents significant challenges. First, Unlike the domains mentioned above, the vast majority of domain knowledge lacks long-term maintained knowledge architectures and publicly available corresponding datasets. This results in existing datasets often being incomplete and unevenly distributed, making it challenging to train a new domain-specific LLM. Second, most domains lack a community of experts. Organizing a large group of experts to define do-

main knowledge architectures and evaluate domain-specific models requires significant human and financial resources, which can significantly slow down the process of domain knowledge construction. Moreover, although in the current work, OceanBench (Bi et al., 2023) introduces the LLMs as an auxiliary tool to reduce the reliance on expert involvement, causing its benchmark evaluation system to primarily focus on natural language capabilities such as analysis, judgment, and classification, rather than the domain-specific knowledge itself.

To address the challenges above, we propose **SDBench**, a generic framework for **constructing domain-specific LLM benchmarks** that leverages survey papers as a substitute for expert involvement. This method is equally applicable to building instruction datasets. Domain surveys, meticulously curated and synthesized by domain experts, are characterized by their comprehensive coverage and timely updates, ensuring both depth and relevance in their respective domains. By using these surveys in combination with LLMs, we have almost replaced the need for extensive expert involvement. Furthermore, the benchmark established through surveys can effectively alleviate the data distribution imbalance issues present in previous works.

To validate the effectiveness and reliability of our approach, we introduce a case study: **BridgeBench** and **BridgeGPT**, the first benchmark and LLM specifically designed for bridge engineering. Bridge engineering, one of the most complex domains in civil engineering, requires dedicated professional expertise for the construction, maintenance, and operation of bridges. The knowledge architecture of bridge engineering is notably extensive and complex, making it challenging for any single expert to comprehend all aspects of the complete lifecycle of a bridge. Through evaluation using **BridgeBench**, we not only demonstrated the domain expertise in bridge engineering tasks of **BridgeGPT** but also validated the efficacy of our entire methodological pipeline.

The contributions can be summarized as follows:

- We propose **SDBench**, a novel generic framework for training and evaluating domain-specific LLMs. This methodology significantly reduces the dependency on domain experts and data engineering resources while maintaining cross-domain applicability and scalability.
- We introduce **BridgeGPT**, the first LLM

specifically designed for bridge engineering. The model demonstrates exceptional domain expertise by handling bridge professional knowledge queries.

- We conduct comprehensive experiments validating the complete methodological pipeline from **SDBench** to construct **BridgeBench** and **BridgeGPT**.

2 Related Work

LLM Benchmarking. In recent years, researchers have focused on quantifying the various capabilities of LLMs. The evaluation methods can be categorized into two main approaches. (1) New algorithms and metrics. Yin et al. (2024); Zhong et al. (2024); Wang et al. (2024c); Dugan et al. (2024) have proposed novel methodologies to assess the knowledge boundaries of LLMs, enabling more accurate judgments of the text generated by these models. Zhang et al. (2024b); Fan et al. (2024); Hashemi et al. (2024); Wang et al. (2024b) have focused on developing methods to assess the reasoning and generative capabilities of LLMs. (2) Introduce domain-specific evaluation datasets. Li et al. (2024b); He et al. (2024); Krumdick et al. (2024); Li et al. (2024a); Chen et al. (2024b,a); Zheng et al. (2024) involves the creation of specialized datasets such as math, finance, news and other specific domains. Zhang et al. (2024a) designed a dataset specifically to test the ability of LLMs to handle ambiguity. Bai et al. (2024) introduced a dataset focused on long-text comprehension. Ren et al. (2024) developed a dataset to evaluate the value alignment of LLMs. SDBench proposed in this paper also falls within this category.

Domain-specific LLM. There have been several related works on domain-specific large models based on professional knowledge, as mentioned in the introduction. Examples of such models include: BioGPT (Luo et al., 2022), designed for solving biomedical problems; ProtLLM (Zhuo et al., 2024), used for protein prediction and generation; MedGPT (Kraljevic et al., 2021), capable of predicting and generating potential medical events; ChiMed-GPT (Tian et al., 2023), trained on traditional Chinese medicine contexts to provide accurate medical diagnostic solutions; GeoGPT (Zhang et al., 2023), utilized for analyzing geographical data; K2 (Deng et al., 2024), a large-scale geographical knowledge model trained on the GeoSignal database; and OceanGPT (Bi et al., 2023),

trained on ocean knowledge bases to address ocean-related tasks. Unlike previous work, we addressed management demands from a real-world bridge and trained **BridgeGPT**, the first LLM specifically designed for the bridge engineering domain, based on a comprehensive bridge corpus.

3 SDBench

In this section, we present the development process of SDBench. With input from only 1-2 domain experts to define the scope of the domain and survey relevant literature, this method can automatically generate a domain knowledge architecture and a benchmark dataset (Section 3.1). Furthermore, by repeating the dataset generation process (Section 3.2), it can also construct an instruction dataset tailored to the domain.

3.1 Survey-based Domain Architecture Construction

This subsection will generate a detailed, multi-level architecture for domain knowledge. As illustrated in Figure 1, we have developed a domain knowledge architecture through the following four steps:

1. *Establish Basic Outline.* The expert defines a domain knowledge outline consisting of first-level headings, with the number limited to no more than ten, and identifies recent survey papers based on the knowledge outline.
2. *Enrich Knowledge Outline.* Using an LLM to generate sub-level headings based on the initial outline from the first step, thereby expanding the domain knowledge architecture.
3. *Survey-based Architecture Update.* First, utilize an LLM to summarize survey papers and generate a survey outline. Then, selectively integrate these survey outlines into the domain knowledge architecture using the LLM.
4. *Domain Expert Investigation.* Use an LLM to determine whether the sections of the survey outline fall within the scope of the parent headings. If a section does not align with the parent heading, it is excluded. This step can optionally incorporate expert review for quality control.

This method is analogous to the construction of knowledge graphs in prior work. By using the capabilities of survey papers, we significantly reduce

manual effort, achieving an almost fully automated process for building knowledge architectures.

3.2 Establishment of Benchmark Dataset

This subsection will ultimately generate a uniformly distributed, domain-specific benchmark dataset. To construct the high-quality benchmark dataset, we follow the methodology outlined in OceanGPT (Bi et al., 2023) and K2 (Deng et al., 2024), utilizing a LLM to extract question-answer (QA) pairs from the domain-specific corpus. Whereas, we have optimized the data cleaning and augmentation steps by introducing an additional data classification phase, as illustrated in Figure 1.

1. *Generate QA Pairs.* Use an LLM (offline LLM for private data) to generate QA pairs that align with the content of the base domain corpus.
2. *Classify QA Pairs.* Employ the LLM to classify the QA pairs into the corresponding lowest-level subheadings of the domain knowledge architecture. This classification can be performed iteratively based on the heading levels.
3. *Decompose QA Pairs.* Use the LLM to decompose QA pairs into true/false questions and multiple-choice questions.
4. *Review and Expand Data.* Review the data volume under each lowest-level subheading. Perform data augmentation for subheadings with insufficient data. This step ensures that the data is uniformly distributed across all categories
5. *Investigate and Refine the Dataset.* Conduct an investigation of the dataset using an LLM. Additionally, expert sampling can be optionally employed to control the overall quality of the dataset, though this step is not mandatory.

By repeating the above steps, a domain-specific benchmark dataset can be rapidly generated. If this method is applied to a larger domain corpus, the output dataset can also serve as an instruction dataset for training domain-specific models. If an instruction dataset needs to be generated simultaneously, the dataset must be partitioned to prevent overlap with the benchmark dataset.

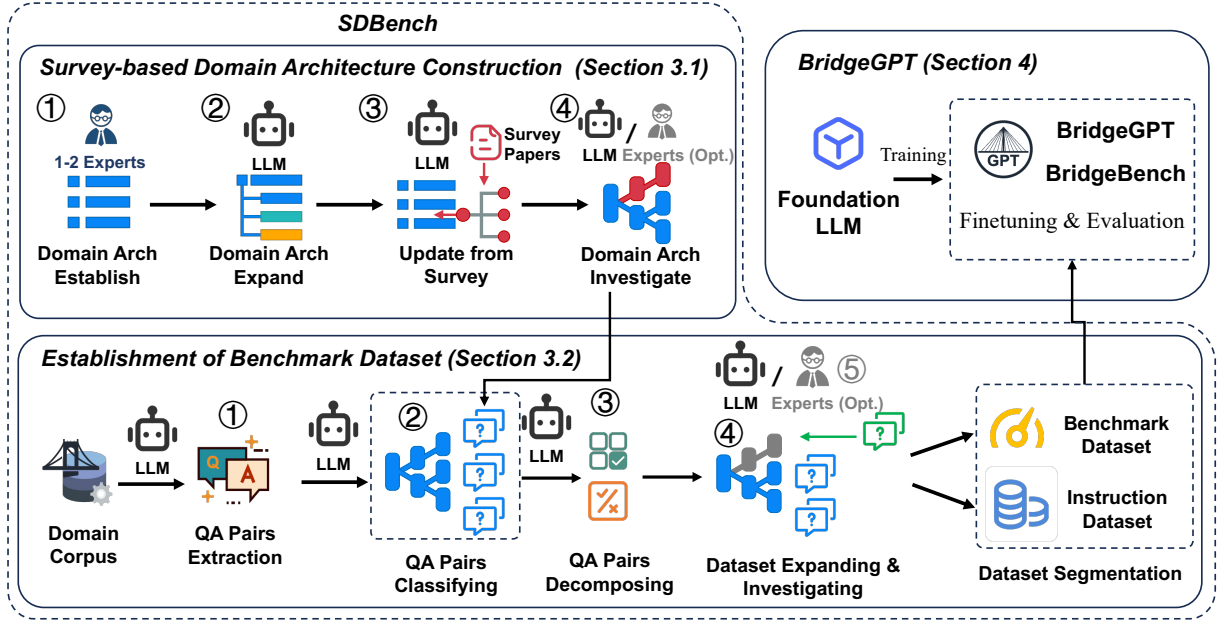


Figure 1: Overview of **SDBench**. For the top-left, an LLM is employed to integrate current domain surveys, enabling the rapid construction of a domain knowledge architecture with minimal expert manpower. Below, the LLM is utilized to systematically generate a domain benchmark dataset based on the established domain knowledge architecture. This method can also be applied to generate instruction datasets. For the top-right, the datasets generated by SDBench can be utilized for domain-specific LLM training and evaluation.

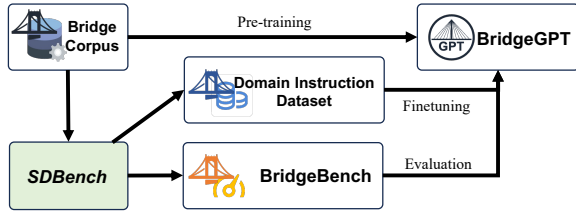


Figure 2: Overview of **BridgeGPT**.

4 Case Study: BridgeBench and BridgeGPT

This section introduces BridgeBench and BridgeGPT as a case study, systematically implementing the SDBench methodology from benchmark design to model construction. Through this comprehensive implementation, we aim to validate the feasibility and reliability of the SDBench approach.

The development of BridgeGPT was motivated by practical engineering requirements from a real-world sea-crossing bridge infrastructure. Given the increasing complexity of bridge management and human resource constraints, the bridge administration authorities sought to use LLMs to optimize core operational processes, including personnel training, report analysis, and operational assessment.

As illustrated in Figure 2, the development of BridgeGPT follows a systematic multi-stage pipeline: (1) construction of a specialized bridge engineering knowledge corpus; (2) development of the BridgeBench dataset via the SDBench; (3) model training (including continual pre-training and fine-tuning) and evaluation.

4.1 Bridge Corpus

The basic task of developing a domain-specific LLM is establishing a high-quality domain corpus. For the bridge engineering domain, we constructed a diverse corpus from multiple sources: primarily 40,070 open-access publications from civil engineering, comprising 38,700 English articles and 1,370 Chinese articles. This was supplemented with publicly accessible web resources, including Bridge-related Wikipedia entries. Additionally, to meet specific engineering requirements, we incorporated private data from the target sea-crossing bridge project, including engineering standards and maintenance knowledge bases.

We employed MinerU (Wang et al., 2024a) to convert various resources into plain text format. To ensure data quality, based on prior experience, we filter out non-essential elements such as author information, headers, footers, and references, while also removing empty lines and other non-textual



Figure 3: An example for knowledge architecture update by survey.

features. Notably, due to the significant knowledge overlap between bridge engineering and civil engineering, the bridge corpus incorporates relevant civil engineering knowledge. This comprehensive corpus was subsequently utilized for the pre-training stage.

4.2 Bridge Knowledge Architecture

Following the methodology outlined in Section 3.1, we constructed a knowledge architecture for the bridge engineering domain. We entrust only one expert to identify four major categories of bridge engineering: bridge exploration, bridge design, bridge construction, and bridge maintenance based on the full lifecycle of bridges. And, a collection of 50 bridge survey papers in bridge engineering has been curated by the expert. Each survey, processed by MinerU, is summarized in an outline by an LLM.

To build the initial knowledge architecture, we first prompted an LLM to generate the structure without inputting any survey data. Then, we iteratively refined and expanded the architecture by incorporating a survey outline from the selected survey papers. Finally, we have generated a knowledge architecture for bridge engineering, consisting of 4 first-level headings, 12 second-level headings, 41 third-level headings, and 272 fourth-level headings. This architecture has been reviewed by a panel of experts and is deemed reasonable by bridge engineering specialists. The detailed knowledge framework is available in our code repository. The prompt used for generating the knowledge architecture is provided in Appendix Table 4.

A demonstration of integrating a survey into the knowledge architecture is illustrated in Figure 3. We use this paper (Sun et al., 2020) as an example, which is a survey discussing the application of deep learning methods in the domain of bridge maintenance. For a simplified architecture establishment

algorithmic workflow, please refer to Algorithm 1. Through the SDBench, it is summarized and inserted into the overall architecture as a three-level subheading.

Algorithm 1

Survey-based Architecture Construction for Bridge

Input: Survey Paper Set $\mathbf{S} = \{S_1, S_2, \dots, S_n\}$

Output: Domain-specific knowledge architecture $\mathbf{D} = \{D_1, D_{1.1}, \dots, D_{p.q.h.k}\}$

- 1: $Expert \rightarrow \mathbf{D}_{init} = \{D_1, D_2, D_3, D_4\}$
- 2: $LLM \rightarrow \mathbf{D}_{extra} + \mathbf{D}_{init} \rightarrow \mathbf{D}'$
- 3: **while** $S_i = True$ **do**
- 4: $LLM \rightarrow S_i \rightarrow S'_i$ ▷ where S' means outline of the survey paper.
- 5: $S'_i \rightarrow LLM \rightarrow \mathbf{D}_{extra}$
- 6: $\mathbf{D}_{extra} \rightarrow LLM \rightarrow \mathbf{D}'$
- 7: **end while**
- 8: $\mathbf{D} \leftarrow \mathbf{D}'$
- 9: **return** $\mathbf{D}$

4.3 BridgeBench

We constructed BridgeBench following the methodology introduced in Section 3.2. Before inputting the basic bridge corpus into the LLM, we performed text segmentation. Through experimentation, we determined that a segment length of 2,000 tokens is optimal for enabling the LLM to extract more QA pairs. Shorter text may result in semantically incomplete, while longer text may lead to information being overlooked by the LLM. Then, these segmented paragraphs were fed into an LLM to summarize and generate QA pairs (offline model for private datasets).

Next, we classified the generated questions to ensure they fall within the bridge engineering domain and assigned them to one of four subcategories. After classification, we further diversified the QA pairs by transforming them into 3-5 multiple-choice and true/false questions using an LLM (offline model for private datasets).

Finally, after a comprehensive review by an LLM, we obtain the BridgeBench dataset. This review ensured the compliance of the content and assessed the distribution of the evaluation dataset. We specifically supplemented the categories that have fewer data samples to ensure a balanced and uniform distribution across the dataset.

As mentioned earlier, this method can also construct an instruction dataset for fine-tuning. We finally established a dataset with around 150,000

Category	Objective	Subjective
Bridge Exploration	2567	690
Bridge Design	2747	780
Bridge Construction	2143	620
Bridge Maintenance	2964	630

Table 1: **BridgeBench** samples for each first heading.

samples. And, currently, we selected more than 13,000 samples (around 8% of the total dataset) to serve as the BridgeBench dataset. Each fourth-level category contains an average of 16 questions, comprising 4 subjective and 12 objective questions. In the current version, 75% of the questions are in Chinese, while the remaining 25% are in English. The ratio of Chinese to English questions can be adjusted based on specific requirements. Data distribution of the second-level heading is shown in Table 1. The algorithmic workflow for establishing BridgeBench can be referred to in Algorithm 2. The prompts used in this section are detailed in Appendix Table 4.

Algorithm 2 BridgeBench Dataset Establishment

Input: Bridge Corpus C ,
Domain-specific knowledge architecture $\mathbf{D} = \{D_1, D_{1.1}, \dots, D_{p.q.h.k}\}$
Output: Benchmark Dataset B , Instruction Dataset I

- 1: Segmentation: $C \rightarrow \mathbf{C} = \{C_1, C_2, \dots, C_n\}$
- 2: **while** $C_i = \text{True}$ **do**
- 3: Extraction: $C_i \rightarrow LLM \rightarrow QA_i$
- 4: Classification: $QA_i \rightarrow LLM \rightarrow D_{p.q.h.k}$
- 5: Decomposition: $QA_i \rightarrow Obj_i \quad \triangleright$ where Obj_i means objective question.
- 6: **if** $D_{p.q.h.k} < tmp$ **then**
- 7: Expending: $LLM \rightarrow QA_i \rightarrow D_{p.q.h.k}$
- 8: **end if**
- 9: **end while**
- 10: $B \leftarrow D * 8\%$
- 11: $I \leftarrow D - B$
- 12: **return** B and I

5 Experiment

5.1 Implementation Details

Implementation of SDBench via BridgeBench: For the online LLM components in **SDBench**, we employed Deepseek-V3, while Qwen2.5-72B/14B-Instruct served as our offline models.

Implementation of BridgeGPT: We selected Qwen2.5-7B-Instruct as the training baseline. We employ Qwen2.5-7B-Instruct and Llama-8B-Instruct as comparative models for BridgeGPT. For the continual pre-training stage and fine-tuning stage, we trained BridgeGPT with four NVIDIA TESLA A800 80G GPUs, $1e-6$ learning rate, and 5 epochs. The fine-tuning method is LoRA (Hu et al., 2021).

Human power of BridgeBench: For the entire development of BridgeGPT and BridgeBench, we only involved two experts from the bridge engineering domain: a Ph.D. and a professor. The Ph.D. primarily focused on collecting and reviewing surveys to construct the knowledge framework, while the professor established the initial knowledge domain classification and provided invaluable advice for BridgeGPT.

Cost of BridgeGPT: The construction of BridgeBench incurred a cost of approximately 2 million tokens, while the fine-tuning dataset for BridgeGPT required around 18 million tokens.

Efficiency of SDBench: The efficiency and automation costs associated with SDBench are influenced by the choice of LLM agent. For instance, utilizing Deepseek-V3, as in this study, with a single-concurrent agent API rate of 20-50 tokens per second, the generation of the BridgeGPT fine-tuning dataset was completed in about 15 days.

5.2 Result of SDBench via BridgeBench

Overall result for BridgeBench. Table 2 presents the accuracy of objective questions in BridgeBench, across various publicly available models. The observed accuracy trends align with the general-purpose capabilities described in the technical reports of these models, demonstrating that BridgeBench, generated by SDBench, possesses a strong reference value. At the 32B and 70B levels, Qwen2.5-32B achieves the highest performance. According to the Qwen2.5 Technical Report (Yang et al., 2024), the 32B version exhibits superior comprehension capabilities compared to the 72B version, giving it an advantage in answering objective questions. Among commercial models, GPT-4o delivers the best results, which is consistent with practical usage experiences. Claude3.5-Sonnet performs between GPT-4o and GPT-4, falling within a reasonable range. We speculate that its corpus in the bridge engineering domain may be slightly smaller than that of GPT-4o.

Commercial Model	(1) Acc.	(2) Acc.	(3) Acc.	(4) Acc.	Total Acc.
GPT-3.5	89.471%	91.65%	89.448%	90.149%	90.18%
GPT-4	90.254%	91.748%	91.629%	91.788%	91.35%
GPT-4o	94.787%	93.784%	95.657%	95.674%	94.98%
Claude3.5-Sonnet	93.127%	92.144%	92.147%	93.21%	92.66%
Open-Source Model					
DeepSeek-R1-Distill-Llama-8B	89.647%	91.245%	90.647%	91.146%	90.67%
Qwen2.5-32B	95.67%	93.805%	94.795%	95.165%	94.86%
DeepSeek-R1-Distill-Qwen-32B	93.784%	92.397%	94.856%	94.632%	93.92%
Qwen2.5-72B	94.155%	94.175%	92.37%	94.085%	93.70%
Llama3-70B	87.64%	89.657%	90.784%	89.147%	89.31%
Qwen2.5-7B	90.398%	91.267%	89.471%	90.746%	90.47%
Llama3-8B	91.76%	89.688%	90.524%	88.269%	90.06%
BridgeGPT-7B (Ours)	92.374%	91.266%	93.787%	93.774%	92.80%

Table 2: The accuracy of various LLMs in answering objective questions from each subset of **BridgeBench**. (1) Bridge Exploration, (2) Bridge Design, (3) Bridge Construction, (4) Bridge Maintenance.

Open-source Models	Acc.
Qwen2.5-7B	67.50%
Qwen2.5-32B	71.20%
Qwen2.5-72B	75.70%
Llama3-8B	62.71%
Llama3-70B	71.20%
DeepSeek-R1-Distill-Llama-8B	59.59%
DeepSeek-R1-Distill-Qwen-32B	62.50%
Commercial Models	Acc.
GPT-3.5	70.70%
GPT-4	74.70%
GPT-4o	80.90%
Claude3.5 Sonnet	76.21%
DeepSeek-V3	83.65%
Ours	Acc.
BridgeGPT-7B	74.80%

Table 3: Comparison of performance among different models on challenging questions subset in bridge engineering.

Discussion for first-level headings for SD-Bench. In Section 3.1, we limit the first-level headings to no more than ten, which is based on our study of well-structured knowledge domains. We found that effectively organized domain knowledge frameworks typically maintain a limited number of first-level headings. For example, Wikipedia’s mathematics section ([Wikipedia contributors, 2025](#)) contains 8 first-level headings. More complex knowledge domains typically expand through second-level headings, which we did not restrict in our approach.

Discussion for BridgeBench Consistency. We

first evaluate the consistency of BridgeBench under the fourth heading classification. To ensure a thorough assessment, human experts were fully familiarized with BridgeBench, and 10% of the questions were randomly sampled for expert evaluation. The experts identified potential errors between the samples and their classifications. The final Inter-Annotator Agreement (IAA) score was 0.8421, demonstrating the dataset’s consistency and validating the effectiveness of the knowledge architecture in BridgeBench.

Additionally, we tested the IAA scores across different classification levels. The IAA scores fell below 0.5 for classifications restricted to either first or second-level headings, as well as for fifth-level heading classifications, making them inadequate as definitive classification standards. For instance, we observed an issue with the question "*Seismic Design of Bridges*", which could belong to both "Bridge Design" and "Bridge Construction." If the classification criteria are too high-level, misclassifications can easily occur, leading to insufficient consistency. On the other hand, overly granular subheadings struggle to cover the entire knowledge scope effectively. Therefore, we decided to divide the bridge knowledge architecture into four levels.

Discussion for the Difficulty of BridgeBench. BridgeBench serves as a comprehensive evaluation dataset, encompassing extensive bridge domain knowledge and covering a wide spectrum of question-answering difficulty levels. We consider that benchmarks with a certain level of inherent difficulty are more effective in significantly differ-

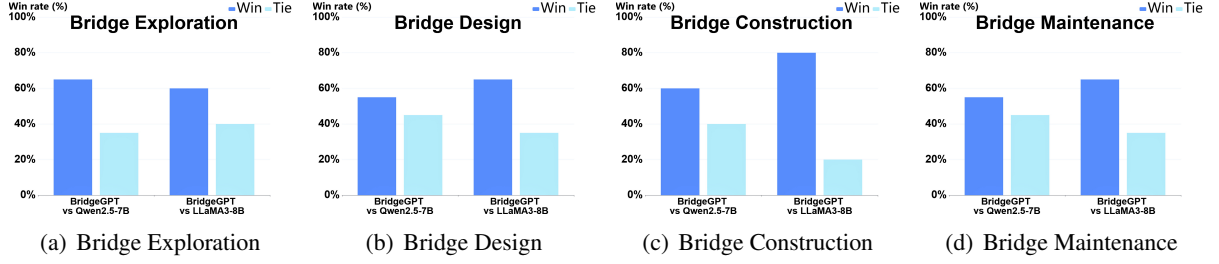


Figure 4: Evaluation result of **BridgeGPT** in the subjective questions answering of **BridgeBench**.

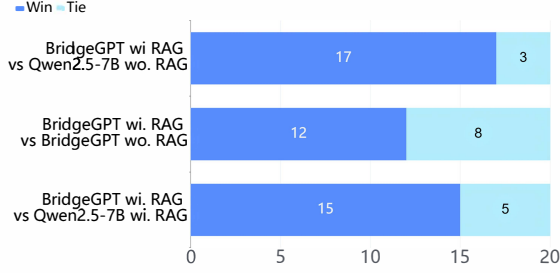


Figure 5: Evaluation result for **BridgeGPT** wi./wo. RAG, competed with Qwen2.5-7B-Instruct.

entiating model capabilities.

To this end, a **challenging** subset, including 60 multiple-choice questions and 40 true/false questions, is curated to specifically evaluate challenging problems within this field. This subset is based on topics highlighted as particularly complex in book (Chen and Duan, 2014), namely bridge loading and structural mechanics.

The results of this subset are shown in Table. 3 which reveals that challenging questions effectively differentiate model capabilities, and BridgeGPT improves performance by 7.3% and 12.09% compared to the same-sized Qwen2.5-7B and Llama3-8B, respectively. The above results demonstrate that our method also shows significant effectiveness on complex domain-specific problems.

5.3 Result of BridgeGPT

Overall result for BridgeGPT. Table.2 demonstrates the performance of BridgeGPT in answering objective questions. As shown, BridgeGPT exhibits a clear advantage over its base model, Qwen-7B-Instruct, when addressing bridge engineering-related problems. For evaluating subjective question responses, we adopt the evaluation methodology for subjective tasks proposed in the OceanGPT (Bi et al., 2023), utilizing an LLM (we employ Deepseek-V3 as the evaluator) to assess the quality of answers generated by different mod-

els. *Win* indicates that a model’s response demonstrates a clear advantage in domain-specific expertise, while *Tie* signifies that there is no significant difference between the models’ responses, with each exhibiting its own strengths. In Figure 4, BridgeGPT demonstrates a significant advantage over Qwen2.5-7B in answering subjective questions, outperforming it across each four knowledge categories of bridge engineering. Additionally, BridgeGPT maintains its superiority over Llama. BridgeGPT demonstrates strong expertise in responding to bridge engineering queries, with detailed answers provided in the Appendix Table 5. Output answers from Qwen2.5-7B and Llama3-8B are provided in the Appendix Table 7.

Evaluation and Discussion for RAG. We also evaluated the impact of RAG on BridgeGPT. For this experiment, we prepared a RAG dataset containing private data related to bridge maintenance and selected 30 subjective questions based on this dataset.

Figure 5 compares the performance of BridgeGPT with and without RAG against Qwen2.5-7B-Instruct. The results demonstrate that BridgeGPT, when combined with RAG, provides highly professional and accurate solutions to bridge engineering problems, showcasing significant potential for engineering applications. Detailed answers of BridgeGPT with RAG are provided in the Appendix Table 6.

5.4 Future Work

During the internal deployment and application, the BridgeGPT has been fully tested by bridge engineering professionals, who have also provided valuable suggestions for improvement. Based on their feedback, we plan to further refine both SD-Bench and BridgeGPT in subsequent work.

First, with the recent open-source release of high-quality LLMs, we will refine the BridgeGPT based on the larger and more intelligent LLMs, particu-

larly in enabling deeper reasoning and analytical thinking.

Second, integrating the multimodal ability into BridgeGPT will be a key focus of our future development efforts, including sensor data from devices such as displacement meters, IMUs (Inertial Measurement Units), and GNSS (Global Navigation Satellite Systems), as well as multimedia data such as videos, audio recordings, and images during bridge maintenance and operations.

6 Conclusion

In this paper, we introduce **SDBench**, a framework for generating benchmark datasets for domain-specific LLMs. This method can also be used to generate instruction datasets for LLM training. To validate the effectiveness of this framework, we present the **BridgeBench**, a novel benchmark for bridge engineering knowledge, and the **BridgeGPT**, the first LLM specialized in bridge engineering, which serves as an expert for bridge engineering tasks. Experimental results demonstrate that this framework enables the construction of high-quality, uniformly distributed training data for domain-specific knowledge and the establishment of a reliable, consistent, and robust evaluation system. Furthermore, BridgeGPT shows high-level expertise in the bridge engineering domain and exhibits the ability to solve practical engineering problems. We will open-source BridgeGPT in the future.

Limitation

SDBench has addressed several critical issues present in prior work, such as data distribution imbalance and uneven data coverage across the training and evaluation stages, while significantly reducing the reliance on expert labor. However, SDBench still lacks the capability to automatically control the quality of domain-specific corpus used for the pre-training stage, which currently requires expert intervention for quality assurance. Additionally, SDBench does not support multimodal evaluation capabilities, which significantly limits the scope of bridge engineering problems that BridgeGPT can effectively address.

For BridgeGPT, the first version was tailored to address the engineering needs of the target administration, with a focus on enhancing its Chinese response capabilities during fine-tuning. We plan to enhance its English response capabilities in fu-

ture versions.

Acknowledgments

This work is supported by Terminus Group, with computational resources and valuable expertise during LLM training provided by Prof. Xian-Sheng Hua and Dr. Shuxian Liang.

This work is supported by the National Key R&D Program of China under Grant No. 2019YFB1600700, the National Natural Science Foundation of China under grant No. 62272407, "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province under grant No. 2023C01033, and the National & Zhejiang Provincial Youth Talent Support Program. Yi Gao and Wei Dong are the corresponding authors.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [LongBench: A bilingual, multi-task benchmark for long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.
- Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2023. OceanGPT: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Wai-Fah Chen and Lian Duan. 2014. *Bridge engineering handbook: construction and maintenance*, volume 5. CRC press.
- Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. 2024a. [Dr.Academy: A benchmark for evaluating questioning capability in education for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

- pages 3138–3167, Bangkok, Thailand. Association for Computational Linguistics.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. [ToMBench: Benchmarking theory of mind in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.
- Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. [RAID: A shared benchmark for robust evaluation of machine-generated text detectors](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.
- Lizhou Fan, Wenye Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2024. [NPHardEval: Dynamic benchmark on reasoning ability of large language models via complexity classes](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4092–4114, Bangkok, Thailand. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. [LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.
- Michael Krumbick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. [BizBench: A quantitative reasoning benchmark for business and finance](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Miao Li, Ming-Bin Chen, Bo Tang, Shengbin Hou, Shengbin Hou, Pengyu Wang, Haiying Deng, Zhiyu Li, Feiyu Xiong, Keming Mao, Cheng Peng, and Yi Luo. 2024a. [NewsBench: A systematic evaluation framework for assessing editorial capabilities of large language models in Chinese journalism](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9993–10014, Bangkok, Thailand. Association for Computational Linguistics.
- Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024b. [GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.
- Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. [ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040, Bangkok, Thailand. Association for Computational Linguistics.
- Limin Sun, Zhiqiang Shang, Ye Xia, Sutanu Bhowmick, and Satish Nagarajaiah. 2020. Review of bridge structural health monitoring aided by big data and

- artificial intelligence: From condition assessment to damage detection. *Journal of Structural Engineering*, 146(5):04020073.
- Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *arXiv preprint arXiv:2311.06025*.
- 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*.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. [Mineru: An open-source solution for precise document content extraction](#). *Preprint*, arXiv:2409.18839.
- Kexin Wang, Nils Reimers, and Iryna Gurevych. 2024b. [DAPR: A benchmark on document-aware passage retrieval](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4313–4330, Bangkok, Thailand. Association for Computational Linguistics.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. [M4GT-bench: Evaluation benchmark for black-box machine-generated text detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.
- Wikipedia contributors. 2025. [Mathematics — Wikipedia, the free encyclopedia](#). [Online; accessed 30-May-2025].
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. [Benchmarking knowledge boundary for large language models: A different perspective on model evaluation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.
- Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024a. [CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.
- Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. 2023. Geogpt: understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*.
- Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024b. [Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1588–1606, Bangkok, Thailand. Association for Computational Linguistics.
- Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. [NEO-BENCH: Evaluating robustness of large language models with neologisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.
- Tianqi Zhong, Zhaoyi Li, Quan Wang, Linqi Song, Ying Wei, Defu Lian, and Zhendong Mao. 2024. [Benchmarking and improving compositional generalization of multi-aspect controllable text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6486–6517, Bangkok, Thailand. Association for Computational Linguistics.
- Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Jianan Zhao, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. 2024. [ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8950–8963, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

Prompt for QA pairs generation:

Assuming you are an expert in the field of bridge engineering, please generate a QA pair based on the provided text, with the following format requirements:

Q: What are the main factors that affect the selection of bridge pile foundation construction?

A: The main factors that affect the selection of bridge pile foundation construction include the following aspects: firstly, geological conditions, the geological structure, soil properties, groundwater conditions, etc. of the construction area will directly affect the form and construction difficulty of the pile foundation. Secondly, the structural requirements of the bridge, including load size, structural type, and span, will put forward different requirements for the design and construction of the pile foundation. Thirdly, for construction equipment and technology, the construction party needs to choose suitable pile foundation construction technology according to the site conditions and construction technical capabilities, such as bored pile, pile sinking method, etc.

Prompt for Generating objective questions through QA pairs:

You are an expert in the field of bridge engineering. Please generate multiple-choice and true or false questions based on the provided QA. Each question should have a corresponding answer. After generation, please check whether these generated questions are correct. If they are not correct, please make modifications.

Prompt for QA classification:

You are an expert in the field of bridge engineering. Now I have some QA pairs that need to be analyzed and classified. Please check if the QA pairs provided by the user match the directory. If they match, please output "match" directly. If they do not match, please give your own suggestions.

Prompt for automatic evaluation using DeepSeek:

Please check if the following sentences contain rich bridge-related information. If so, output "related". Otherwise, output "unrelated".

Sentences: With the development of the Internet of Things, artificial intelligence, and big data analysis, the health monitoring of bridge structures will increasingly adopt intelligent materials and technologies. For example, sensors embedded in smart concrete or steel can monitor the structural status in real-time.

Table 4: Prompt for Generate objective questions through QA pairs of **SDBench**

Instruction:	What are the system functions of the integrated facilities of the Hong Kong Zhuhai Macao Bridge system?
BridgeGPT:	1.Data collection: - Structural subsystem: Real-time monitoring of the structural health status of the bridge through various sensors (such as stress sensors, strain sensors, etc.)- Traffic flow subsystem: collects data on vehicle flow, speed, and driving paths to provide a basis for traffic management and optimization. 2. Information exchange: - Data transmission: Real-time transmission of data from various subsystems to the data center through high-speed networks, achieving centralized management and analysis of data- Information sharing: Ensure timely transmission and sharing of information between subsystems to improve decision-making efficiency- Collaborative work: With the support of system integration facilities, different subsystems can work together to respond to unexpected situations. 3. Internet of Things applications: Bridge monitoring system: Through Internet of Things technology, real-time monitoring of stress, strain, and other key parameters of bridges can be achieved, and potential problems can be detected and warned in a timely manner- Fatigue risk assessment: Based on historical data and current environmental conditions, evaluate the fatigue life of bridges under different loads and develop corresponding maintenance plans- Health Management System: Utilizing big data analysis technology to comprehensively monitor and manage the health of bridges, ensuring their long-term stable operation. 4. Vehicle driving control: - Intelligent navigation system: provides real-time navigation services for drivers, optimizes driving paths, and reduces congestion- Speed limit and parking reminder: dynamically adjust speed limits based on traffic flow and environmental conditions, and issue parking reminders when necessary to ensure traffic safety- Emergency response: In the event of a traffic accident or other emergency situation, be able to quickly respond and take measures to ensure personnel safety. 5. Navigation service function: - Real time update of traffic information: Through map software and in the car navigation system, real-time traffic information is provided to help drivers avoid congested roads.

Table 5: Output for **BirdgeGPT**

Instruction:	What are the system functions of the integrated facilities of the Hong Kong Zhuhai Macao Bridge system?
BridgeGPT + RAG:	The system integration has the following main functions: 1. Equipment integration management can monitor and control the operation status of equipment in real-time through the equipment centralized control platform and can perform cross-system linkage control of equipment according to various control plans. It is also possible to manage the maintenance and upkeep of equipment by establishing electronic and full-life archives for various facilities and equipment. The various system facilities included in integrated management include: a monitoring system, ventilation system, fire protection system, water supply and drainage system, power supply and distribution system, lighting system, etc. 2. System redundancy: Computer management systems with servers as the core are established in both the management center and the monitoring station on the west artificial island. In order to improve system performance and reliability, the system integration platform is configured for hot standby in the West Artificial Island cluster, and the management center is also equipped with an integrated platform. The West Artificial Island and the management center form a disaster recovery relationship. When a comprehensive disaster occurs on the West Artificial Island, the management center activates its integrated platform through manual intervention to continue providing services, and the operations of management personnel are not affected. 3. The data storage disaster recovery management center and the data storage system configured for monitoring can achieve real-time synchronization of production data and backup data between the two locations. Under normal circumstances, storing data in two locations simultaneously is like virtualizing a set of storage for upper-level application systems; When a system failure occurs in one location, the other system independently assumes the data storage function. When the failure is restored, the faulty system automatically synchronizes and updates to the latest data. 4. Equipment tag management: Establish a device identification management system based on RFID tags and QR code tags. Management personnel can use handheld devices to read the device information stored in the tags and enter the current device status information. It is also possible to remotely query detailed information about devices and facilities through wireless communication based on the device ID stored in the tag, providing electronic means for equipment inspection and maintenance management. 5. The system interconnection and integration platform has the function of exchanging information with other management information systems. The system can receive information from other information systems, call corresponding plans based on the information content, and send feedback information to the other party. This system can also actively send alerts or queries to other information systems and receive feedback information from them.

Table 6: Output for **BirdgeGPT** with RAG

Instruction:	What are the system functions of the integrated facilities of the Hong Kong Zhuhai Macao Bridge system?
LLaMA3-8B-Instruct:	The system functions of the integrated facilities of the Hong Kong Zhuhai Macao Bridge include: 1. Traffic management system: responsible for managing and controlling the traffic flow of the bridge, including traffic signal control, traffic monitoring, vehicle detection, and other functions. 2. Communication system: Provides efficient communication services, including wireless communication, wired communication, broadcasting, and other functions. 3. Energy Management System: Responsible for managing and controlling the energy usage of the bridge, including functions such as power supply, water supply, and gas supply. 4. Environmental monitoring system: monitors and controls the environmental conditions around the bridge, including air quality, water quality, acoustic environment, and other functions. 5. Traffic Information Release System: Release traffic information, including real-time traffic conditions, road conditions, weather, and other information. 6. Intelligent transportation system: Utilizing data analysis and artificial intelligence technology to optimize traffic flow, and reduce congestion and accidents. 7. Maintenance Management System: Responsible for the maintenance and upkeep of the bridge, including preventive maintenance, emergency repairs, and other functions. 8. Security Protection System: Provides security protection services, including anti-theft, anti-robbery, and other functions. 9. Integrated management system: responsible for integrating and managing all system functions, providing comprehensive and efficient management services.
Qwen2.5-7B-Instruct:	The integrated facilities of the Hong Kong Zhuhai Macao Bridge system are an important component of the bridge, and its system functions mainly include the following aspects: 1. Traffic monitoring system: Through the installation of cameras, sensors and other devices on the bridge, real-time monitoring of the traffic conditions on the bridge deck, including traffic flow, speed, accidents, etc., is carried out for timely traffic dispatch and management. 2. Meteorological monitoring system: Real-time monitoring of meteorological conditions in the bridge area, including wind speed, direction, temperature, humidity, visibility, etc., providing meteorological data support for the safe operation of the bridge. 3. Environmental monitoring system: Monitor the environmental conditions of the bridge area, including water quality, noise, air pollution, etc., to protect the local ecological environment. 4. Emergency rescue system: Provide emergency rescue services, including emergency call systems, rapid response mechanisms for ambulances and fire trucks, etc., to ensure prompt action in case of emergency. 5. Lighting control system: Automatically adjust the lighting intensity and mode of the bridge according to different time periods and weather conditions to ensure driving safety at night or in adverse weather conditions. 6. Bridge health monitoring system: By installing sensors in the bridge structure, real-time monitoring of the structural status of the bridge, including stress, displacement, vibration, etc., ensures the safety and durability of the bridge.

Table 7: Output for LLaMA3-8B-Instruct and Qwen2.5-7B-Instruct