LISGPT: Research on the Construction of a Library and Information Science Academic LLM Based on the Boundary Knowledge Enhance Framework

Zhu, Yu Nanjing University, China | zhu.yu@smail.nju.edu.cn

Duan, Yongkang Beijing Normal University, China | duanyongkang@mail.bnu.edu.cn

Hu, Hengjia China University of Geosciences, China | 2428002915@qq.com

Jin, jian Beijing Normal University, China | jinjian.jay@bnu.edu.cn

Ye, Jiyuan Nanjing University, China | yejiyuan@mail.nju.edu.cn

ABSTRACT

Academic large language models have demonstrated transformative potential in natural language processing tasks. However, they still face significant challenges in adequately understanding highly specialized and complex domain-specific knowledge. To address this issue, this study introduces the Boundary Knowledge Enhance (BKE) framework, which constructs a large-scale, high-quality professional question-answering dataset (n = 276,083) in the Library and Information Science (LIS) domain, specifically designed to capture the complexity of social science knowledge. Furthermore, by employing the proposed Direct Boundary Knowledge Optimization (DBKO) training method, the model's ability to comprehend and apply specialized domain knowledge is significantly enhanced. Experimental results show that LISGPT achieves superior performance compared to state-of-the-art commercial models. In the literature keyword prediction task, it outperforms all baseline models with an F1 Score of 0.3973, ranking first. In the professional translation task, it reaches 99.1% of the performance level of DeepSeek-V3-671b, achieving an average score of 0.5971 and ranking third. Ablation studies confirm that the overall performance improvement of LISGPT after DBKO training is 2.32%. This study open-sources the large LIS training datasets and three versions of a specialized LIS academic model, offering a practical paradigm for developing open-source, efficient models in other humanities and social sciences domains.

KEYWORDS

LISGPT, Academic Large Language Model, Large Language Model, Library and Information Science, Boundary Knowledge Enhance

INTRODUCTION

The large language models (LLMs) represented by GPT-4 have demonstrated significant potential in various natural language processing tasks, including information aggregation, reasoning, and generation. These capabilities have been explored in highly specialized fields such as medicine (Thirunavukarasu et al., 2023), finance (Wang et al., 2023), and law (Zhou et al., 2025). This transformative technology has also sparked considerable interest in the Library and Information Science (LIS) domain. General-purpose LLMs exhibit strong competence in downstream applications such as identifying meeting action items (Sadia et al., 2025), information retrieval (Yu et al., 2024), sentiment analysis (Lu et al., 2025), rumor detection (Chen et al., 2025), and academic text evaluation (Thelwall, 2025), supporting the entire workflow of the discipline from data extraction, analysis, and evaluation to intelligence provision.

Despite initial successes in LIS-related reasoning tasks, the widespread use of LLMs faces significant practical limitations. Closed-source proprietary models like GPT-4o and extremely large open-source models such as DeepSeek V3 require unpredictable API access and deployment costs, introducing substantial data privacy risks and high inference expenses. Research suggests that more specialized, vertically focused, low-resource models can achieve superior performance with better efficiency (Ye et al., 2025). Clearly, obtaining high-quality, relevant, and boundary-specific knowledge-level data is critical for developing effective and efficient open-source LIS-LLMs, addressing the insufficient understanding of LIS knowledge in current general-purpose models, as noted by Dervin (1998), and overcoming barriers in implementing large-scale upstream tasks.

To address these challenges, we propose **B**oundary **K**nowledge **E**nhance (BKE), a framework designed for reasoning tasks in social science domains like LIS. BKE enables our trained model, LISGPT, to achieve performance comparable to commercial, closed-source LLMs, demonstrating its effectiveness. Tailored to the complexity of LIS domain knowledge, we constructed a large-scale, high-quality LIS professional question-answering dataset (n=276,083). We introduced a refined and comprehensive training paradigm that seamlessly integrates the best open-source foundational models with LIS domain-specific knowledge data. Additionally, we open-sourced a vertically specialized academic LLM tailored for upstream LIS applications (URL for LISGPT model weights: https://www.modelscope.cn/models/YKDuan/IRM_chat_3B, https://www.modelscope.cn/models/

YKDuan/IRM_chat_7B, https://www.modelscope.cn/models/YKDuan/IRM_chat_14B URL for LISGPT dataset: https://www.modelscope.cn/datasets/YKDuan/IRM_chat_all). This work provides a feasible paradigm for constructing open-source, efficient academic LLMs in other humanities and social sciences, paving new paths for the development of intelligent academic research support tools.

RELATED WORK

Academic Large Language Models

With the rapid advancement of artificial intelligence, LLMs have demonstrated significant potential across multiple domains. However, current general-purpose LLMs still face challenges in providing adequate knowledge support for academic applications, particularly in the social sciences. This issue stems not only from the intrinsic characteristics of social science knowledge systems—such as the complexity of conceptual relationships, the diversity of theoretical perspectives, and the uncertainty and context-dependency of research conclusions—but also from inherent limitations of LLMs, including hallucinations, biases, and reliance on outdated information. These factors impose higher demands on the knowledge representation capabilities of such models (Cuskley et al., 2024; Grossmann et al., 2023).

In the ecosystem of academic LLMs, the development trajectories of Chinese and English models exhibit distinct characteristics. A representative product in the Chinese domain is Spark Research Assistant, developed through a collaboration between iFlytek and the National Science Library of the Chinese Academy of Sciences. This closed-source commercial model integrates functionalities such as literature review, paper analysis, academic writing, and intelligent research assistance (Qian et al., 2024). However, the lack of transparency regarding its training data, parameter settings, and algorithm implementation limits its utilization and improvement by the academic community. Another notable Chinese academic LLM is Huazhi Wensi (CNKI, 2024), developed by CNKI and Huawei Cloud, which combines pre-training, fine-tuning, and RAG enhancement techniques to support intelligent search, reading, and question-answering functions (Huazhi, 2024). In contrast, the development of English academic LLMs is more diverse. Models like Scholar GPT (CharityGPT, 2024) and Elicit AI (Whitfield & Hofmann, 2023) adopt a technical approach of fine-tuning general-purpose LLMs with academic corpora, excelling in tasks such as literature comprehension, interdisciplinary knowledge integration, and academic writing assistance. Additionally, Clarivate's release of the Web of Science Research Assistant in September 2024 further enriches the English academic LLM ecosystem (Clarivate, 2024). While these academic LLMs collectively advance the development of intelligent research support tools, their commercial nature still results in notable shortcomings in openness and transparency.

Moreover, most mainstream academic LLMs adopt a broad and comprehensive knowledge coverage strategy. For instance, while Spark exhibits extensive cross-disciplinary knowledge, it performs poorly in fine-grained knowledge representation within specific disciplines (Ling et al., 2024). This coarse-grained knowledge structure struggles to meet the rigorous demands of professional academic research for conceptual precision and theoretical detail. Particularly in fields like the social sciences, where the accuracy of concepts is paramount, insufficient knowledge granularity leads to issues such as conceptual confusion and incorrect application of theories when handling specialized academic information processing tasks (Ziems et al., 2024).

Domain-Specific Large Language Models

General-purpose large models face significant challenges in complex upstream natural language processing tasks within the social sciences, particularly in terms of insufficient knowledge granularity and a lack of open-source transparency. These challenges highlight the urgent need for domain-specific large models. The specialization and verticalization of models are key to overcoming the limitations of general-purpose models (Suzuki et al., 2023). By integrating professional knowledge systems and adapting model architectures to fit specific knowledge structures, domain-specific large models can more accurately capture professional concepts, theoretical frameworks, and methodologies, thereby providing more precise intelligent support for deep academic information processing in the social sciences (Wu et al., 2024).

For example, in the financial domain, Lee et al. (2025) analyzed five key technical approaches adopted by eight financial large models, including parameter-efficient fine-tuning, instruction tuning, and enhanced context learning. These models demonstrated superior performance across six benchmark tasks, such as financial text classification, sentiment analysis, and named entity recognition. Xie et al. (2023) introduced the PIXIU framework, which developed the FinMA model based on instruction tuning. Liu et al. (2025) pioneered a new path for multimodal financial models with their AT-FinGPT. In the legal domain, research on large models also exhibits a trend toward diversified technical methods and refined application scenarios. Shi et al. (2024) designed the Legal-LM model using a two-stage training method, integrating legal knowledge graphs with large language models to significantly enhance the model's understanding of complex relationships between legal concepts. Yao et al. (2024) proposed Lawyer GPT, which employs a retrieval-reasoning-validation three-module framework and utilizes recursive chainof-thought reasoning for multi-step inference, achieving accuracy close to that of human legal graduate students. Li et al. (2024) introduced the continuous semantic augmentation fine-tuning method, which adopts a retain-answer-transform-question strategy to ensure the professionalism of generated data while increasing the diversity of question formulations. Remarkably, this method achieved performance comparable to full-data training using only 5% of the original dataset. Clearly, domain-specific large models achieve more precise mastery of domain-specific concepts, theories, and methodologies through the integration of professional knowledge systems and adjustments to model algorithms.

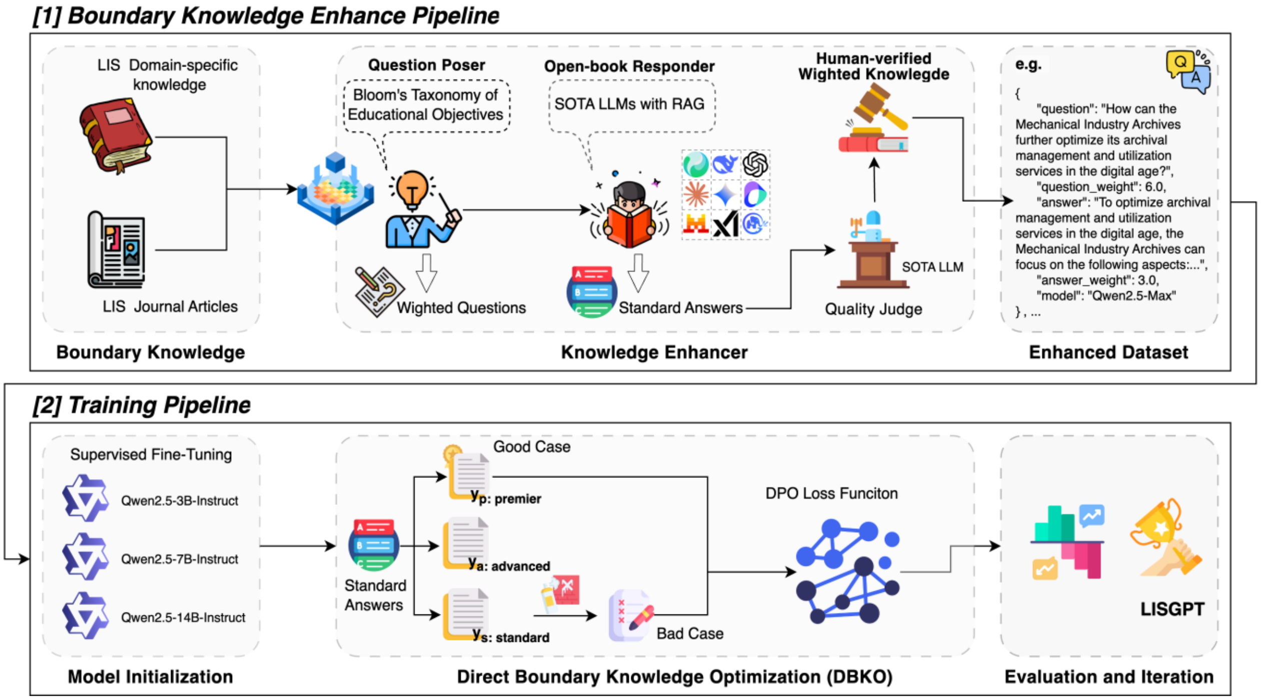
To sum up, in the era of artificial intelligence, LIS academic research faces challenges such as a lack of large-scale, high-quality training data and the complexity of knowledge organization. This highlights the need to explore the development of vertically specialized large models to enhance information processing and knowledge management efficiency. Drawing on experiences from other social science domains, constructing multi-source, high-quality domain knowledge datasets through a boundary knowledge enhancement framework can lead to the creation of domain expert models that understand the semantic relationships of deep LIS domain knowledge while serving upstream LIS information reasoning tasks.

METHODOLOGY

The synthesis of data tailored to the LIS domain poses significant challenges for LLMs due to the following two reasons: (1) The absence of authoritative and in-depth LIS knowledge in the data augmentation and distillation processes of existing state-of-the-art (SOTA) LLMs limits the quality and diversity of synthesized data; (2) The difficulty in formalizing and validating LIS synthetic data makes it highly challenging to evaluate data quality and eliminate hallucinations after knowledge enhancement.

To address these issues, this section details the complete training process of LISGPT, focusing on constructing knowledge-level data to enhance the reasoning performance of open-source LLMs in the LIS domain. We propose the BKE framework based on SOTA LLMs, addressing the first challenge by incorporating authoritative LIS knowledge documents and high-quality journal papers as the knowledge database. To tackle the second challenge, we employ a Knowledge Enhancer to establish progressively sophisticated question generators and answer sets, ensuring all data meets quality standards through human-AI collaboration.

Furthermore, to further improve the reasoning performance of the model in the LIS domain, we implement the $\boxed{\text{D}}$irect $\boxed{\text{B}}$oundary $\boxed{\text{K}}$nowledge $\boxed{\text{O}}$ptimization (DBKO) method based on open-source SOTA models. This approach guides LLMs in the social sciences to enhance their generalization of professional knowledge while maintaining efficient and accurate instruction-following capabilities. An overview of the entire process is illustrated in Figure 1.



```
{
    "question": "How can the Mechanical Industry Archives further optimize its archival management and utilization services in the digital age?",
    "question_weight": 6.0,
    "answer": "To optimize archival management and utilization services in the digital age, the Mechanical Industry Archives can focus on the following aspects:...",
    "answer_weight": 3.0,
    "model": "Qwen2.5-Max"
}, ...
```

Figure 1. Process Framework

Data Construction

Boundary Knowledge Enhance

(1) Boundary Knowledge Database

Let Θ denote the parameters of a given LLM and let K represent the set of all knowledge currently known to humans. For each piece of knowledge $k \in K$, there exists an associated set of question-answer pairs $Q_k = \{(q_i, a_i)\}_i$. The inclusion of a specific piece of knowledge within the boundary of the LLM's knowledge can be evaluated using the conditional probability $P_\Theta(a_i \mid q_i)$ and a given threshold $\epsilon \in (0.5, 1]$ (Yin et al., 2024). K can be further divided into three mutually exclusive subsets:

- Core Knowledge (K_C): Also referred to as Prompt-Agnostic Knowledge (Yin et al., 2024), this subset represents the knowledge embedded in the LLM's parameters Θ, which can provide correct answers regardless of the phrasing of the question. $K_C = \{k \in K \mid \forall(q_i, a_i) \in Q_k, \ P_\Theta(a_i \mid q_i) \rangle \epsilon\}$.
- Unknown Knowledge (K_U): This subset represents the knowledge absent from the LLM's parameters Θ, for which the model cannot provide correct answers under any phrasing of the question. $K_U = \{k \in K \mid \forall(q_i, a_i) \in Q_k, P_\Theta(a_i \mid q_i) < \epsilon\}$.
- Boundary Knowledge (K_E): This subset represents the knowledge present in the LLM's parameters Θ, but only for specific phrasings of the question can the model provide correct answers. $K_E = \{k \in K \mid k \notin K_C \wedge k \notin K_U\}$.

The boundary between K_E and K_U is referred to as the parametric knowledge boundary (M. Li et al., 2024), within which the knowledge is contained in the LLM's parameters Θ. From the perspective of the Sense-Making Theory, this parametric knowledge boundary represents the gap that the model faces when constructing meaningful interpretations in specialized domains. To provide a bridge for LLMs to cross this knowledge boundary (Dervin, 1998) and enable them to construct meaningful responses in the LIS professional context, it is necessary to construct a Boundary Knowledge Database K_B. This knowledge database must satisfy the conditions $K_B \cap K_U \neq \emptyset$ and $K_B \cap K_E \neq \emptyset$ aiming to expand the LLM's K_E in the LIS domain and transform more K_U into comprehensible knowledge for the model.

To ensure the breadth of coverage in the knowledge database, the first type of boundary knowledge K_{B1}, is sourced from authoritative LIS knowledge documents. This involves collecting relevant entries from authoritative encyclopedic websites related to the LIS domain, including professional concepts, theoretical frameworks, and methodologies, forming the core foundational knowledge of the knowledge database. After cleaning, N entries are obtained, with each entry $k_a \in K_{B1}$ consisting of a title t_a and content c_a, represented as: $K_{B1} = \{(t_a, c_a) | a \in [1, N]\}$. To ensure the depth of boundary knowledge in the knowledge database, the second type of boundary knowledge, K_{B2}, is derived from high-quality journal articles in the LIS domain. These articles are primarily sourced from core LIS journals, and after screening, M high-quality documents are obtained. Each document $k_b \in K_{B2}$ includes bibliographic data represented as: $K_{B2} = \{(T_b, E_b, W_b) | b \in [1, M]\}$, where T_b, E_b, and W_b represent the Chinese title, English title, and keyword set of the document, respectively. The combination of these two data sources forms the complete boundary knowledge database: $K_B = K_{B1} \cup K_{B2}$.

(2) Knowledge Enhancer

The boundary knowledge database K_B provides static boundary knowledge. To transform it into high-quality question-answer pairs, we developed a Knowledge Enhancer based on dynamic cognitive principles. This Knowledge Enhancer consists of two modules: the Question Poser, driven by Bloom's Taxonomy (Bloom, 1984), and the Open-book Responder, which leverages retrieval-augmented techniques.

The Question Poser takes the K_B as input and generates multi-level, multi-dimensional questions based on Bloom's Taxonomy of Educational Objectives (Bloom, 1984). Bloom's Taxonomy categorizes cognitive processes into six hierarchical levels — remembering, understanding, applying, analyzing, evaluating, and creating — progressing from lower-order to higher-order thinking. Using this framework, the Question Poser generates questions at different cognitive levels for each knowledge source $k \in K_B$. Each generated question q is assigned a corresponding weight $q^w \in \{1, 2, \ldots, 6\}$, where the weight corresponds to one of the six cognitive levels in Bloom's Taxonomy. A higher weight indicates that the question involves a higher cognitive level and requires deeper knowledge comprehension.

On one hand, the Question Poser leverages a large language model LLM_Q, using each element $k_a \in K_{B1}$ from the first type of boundary knowledge as input to generate multiple questions $q_L$based on Bloom's Taxonomy, along with their corresponding question weights (cognitive levels) q_L^w. By prompting the LLM at multiple cognitive levels, a set of questions $Q_L$is generated for the authoritative LIS knowledge K_{B1}:$Q_L = \left\{(q_L, q_L^w) | q_L = LLM_Q(k_a), k_a \in K_{B1}, q_L^w \in \{1, 2, \ldots, 6\}\right\}$. On the other hand, to fully utilize the boundary knowledge in the knowledge database, the Question Poser employs a predefined set of question templates T and the metadata from K_{B2} to construct structured questions. Each predefined template $t \in T$ is associated with a corresponding

Bloom's Taxonomy weight q_t^w, enabling the generation of questions qt based on $k_b \in K_{B2}$. This process yields a set of questions $q_t$for the high-quality journal paper knowledge K_{B2}: $Q_T = \{(q_t, q_t^w) | q_t \in \{t(k_b) | t \in T\}, k_b \in K_{B2}, q_t^w \in \{1, 2, \ldots, 6\}\}$. Ultimately, the questions generated through both approaches are merged into a unified question set:$Q = Q_L \cup Q_T$.

When humans encounter knowledge gaps, they actively construct bridges by searching for relevant information. To simulate this behavior in meaning-making rather than merely regurgitating preset answers, the Open-book Responder adopts retrieval-augmented generation (RAG) technology, combining SOTA large models LLM_A with the boundary knowledge database K_B. Generating high-quality RAG responses that align with human preferences places high demands on the contextual processing and reasoning capabilities of LLM_A. To identify models that excel in these comprehensive abilities, Chatbot Arena (Chiang et al., 2024) collects large-scale crowdsourced data on real user preferences. Its evaluation results have been shown to align highly with expert judgments. Therefore, top-ranked models on this leaderboard, such as GPT-4o and DeepSeek V3, are considered SOTA in generating high-quality, human-preferred content. Based on this criterion, we selected these SOTA LLMs as LLM_A. For each question $q \in Q$, the Open-book Responder first retrieves the most relevant knowledge snippets from the knowledge database K_B and then uses m selected SOTA large models to generate corresponding standard answers a_i based on these retrieval results: $A_i = \{a | a = LLM_A(q, K_B), q \in Q\}$. Through the Question Poser and Open-book Responder, the Knowledge Enhancer ultimately constructs i Q&A pairs, forming the initial dataset: $D_{init} = \{(q_i, a_i, q_i^w) | (q_i, q_i^w) \in Q, a_i \in A\}$.

Knowledge Validation

Although the initial dataset D_{init} generated by the Knowledge Enhancer is based on authoritative knowledge, it may still suffer from issues such as model hallucinations. To ensure the quality of the training data, we employ a two-step validation process combining preliminary screening by LLMs and final evaluation by human experts.

The Quality Judge module utilizes a SOTA model as the Judge Model (LLM_J). Based on the knowledge database K_B, this module evaluates the multiple answers ai corresponding to each initial question q_i, assessing their quality across three dimensions: accuracy, comprehensiveness, and professionalism. The quality score a_i^w is construct as $a_i^w = LLM_J(q_i, a_i, K_B)$, and the answers are categorized into three quality levels, assigned weights $a_i^w \in \{1, 2, 3\}$, with higher weights indicating superior overall quality across the evaluated dimensions. Subsequently, human experts validate the assigned answer weights a_i^w, adjusting them to reflect their quality assessments. This step ensures that the final dataset reflects both automated and expert evaluations. The resulting validated dataset is represented as:

$$D = \{(q_i, a_i, q_i^w, a_i^w) | i \in [1, N]\}$$

Training Method

Model Initialization

This study selected three pre-trained models of varying scales as the foundation: Qwen2.5-3b-Instruct, Qwen2.5-7b-Instruct, and Qwen2.5-14b-Instruct. The choice of these instruction-tuned models over Base or Chat variants was primarily motivated by their superior capabilities in instruction understanding and execution. Instruction-tuned models undergo specialized training for instruction-following, demonstrating exceptional task recognition and intent-execution precision. This enables them to respond to user queries with higher accuracy (Chung et al., 2024) — a trait that aligns well with the objective of this research to construct a high-quality academic large model tailored for upstream tasks.

These three models were initially fine-tuned using our enhanced dataset in conjunction with Low-Rank Adaptation (LoRA) technology (Hu et al., 2021). This process endowed the models with fundamental domain-specific question-answering capabilities, laying the groundwork for subsequent in-depth optimization.

Direct Boundary Knowledge Optimization

We propose a static preference data-based direct policy optimization method, **D**irect **B**oundary **K**nowledge **O**ptimization（DBKO), aimed at further enhancing the model’s generalization and reasoning capabilities in specialized domain knowledge. By introducing quality-stratified preference boundary knowledge contrastive data, DBKO improves alignment efficiency without requiring explicit reward models or dynamic environment interactions.

(1) Quality Stratification of Training Data

The core of DBKO lies in the rigorous quality stratification of training data. In this study, standard answers are categorized into three classes based on their quality:

a. Premier Answer (y_p): Answers that have been validated by the judge model and scored highest in terms of accuracy, comprehensiveness, and professionalism. These responses are considered good cases.

b. Advanced Answer (y_a): High-quality responses that contain correct information but may lack comprehensiveness or exhibit less professional phrasing.

c. Standard Answer (y_s): Basic correct answers that meet minimum correctness criteria.

To construct contrastive learning samples, this study intentionally introduces malicious alterations (ϕ) to a subset of standard answers, converting them into bad cases. The rewriting process includes introducing inaccurate technical terminology, removing critical qualifying conditions, or inserting content that conflicts with domain-specific knowledge.

(2) Weighted Square Ratio Sampling Method

This study proposes the **W**eighted **S**quare **R**atio **S**ampling (WSRS) method to construct high-quality premier answers and low-quality standard answers as training samples from a large-scale question-answer pair dataset. The core idea of WSRS is to comprehensively utilize information from both the question weight and answer weight dimensions, employing nonlinear mapping and stratified sampling to enhance the overall quality of the training data while ensuring the representativeness of questions across all quality levels.

The design of WSRS is partially inspired by Bradford's Law, which reveals the long-tail characteristics of scientific journal article distributions and effectively delineates the core regions of information resources (Bradford, 1934). By analogy to the question-answering domain, this study finds that the distribution of question weights also exhibits long-tail characteristics. Specifically, Premier questions, though relatively fewer in number, typically contain more valuable knowledge and play a critical role in determining answer quality. In contrast, the numerous standard questions contribute relatively less to knowledge representation. This insight highlights the importance of prioritizing higher-quality questions and answers during sampling while maintaining a balanced representation of lower-quality data.

Based on the above analysis, WSRS employs a nonlinear weight mapping method to assign higher sampling probabilities to premier questions, aiming to maximize the overall knowledge value of the sampled training set. Given the original question-answer dataset $D = \{(q_i, a_i, q_i^w, a_i^w)\}_{i-1}^N$, where q_i is the i-th question, a_i is the corresponding answer, $q_i^w \in \{1, 2, \ldots, 6\}$ is the question weight, and $a_i^w \in \{1, 2, 3\}$ is the answer weight, WSRS first partitions D into k mutually exclusive subsets based on the question weights q_i^w:$D = \bigcup_{j=1}^k D_j$, satisfying $\forall (q, a, q^w, a^w) \in D_j, w = j$. For each subset D_j, WSRS calculates its sampling quantity using the WSRS formula: $p_j = \frac{j^2}{\sum_{l-1}^k l^2} \cdot p$, where p is the total number of samples to be drawn from the dataset D. After determining the sampling proportions for each subset, WSRS further selects question-answer pairs (q_i, a_i, q_i^w, a_i^w) with answer weight $a_i^w = 3$ from each D_j, forming the Premier answer set $A_p = \left\{(q, a, q^w, a^w) | (q, a, q^w, a^w) \in \bigcup_{j=1}^k SampleByRadio\left(D_j, p_j\right) \land s = 3\right\}$, where $SampleByRadio\left(D_j, p_j\right)$ denotes random sampling from D_j at the ratio p_j.

For each question-answer pair $(q, a, q^w, 3)$ in A_p, the corresponding pair $(q, a_s, q^w, 1)$ with answer weight $a^w = 1$ is identified from the original dataset D, forming the standard answer set $A_s = \left\{(q, a_s, q^w, 1) | \left(q, a_p, q^w, 3\right) \in A_p \land (q, a_s, q^w, 1) \in D\right\}$. For each answer as in A_s, we generates its variant $\varphi(a_s)$ through intentional malicious alterations φ, resulting in the enhanced standard answer set $\hat{A}_s = \left\{\left((q, \varphi(a_s), q^w, 1)\right) | (q, a_s, q^w, 1) \in A_s\right\}$. The premier answer set A_p and the enhanced standard answer set $\hat{A}_s$ are then used to construct a boundary domain knowledge preference dataset. DBKO training is performed based on the direct preference optimization (DPO) loss function (Rafailov et al., 2023), thereby enhancing the model's ability to understand and apply specialized domain knowledge.

EXPERIMENTS

Experimental Settings

Comparison Models

We selected both open-source and closed-source SOTA models as benchmarks for comparison, as shown in below: DeepSeek-V3, Qwen2.5-Max, Qwen2.5-14b/7b/3b-Instruct, Llama-3.1-8B-Instruct, ERNIE-Tiny-8K, Doubao-1.5-Lite, Claude-3-Haiku, GPT-4o mini.

Dataset Construction

We utilized SOTA LLMs such as GPT-4o and DeepSeek V3 to implement the DBKO framework, constructing a question-answer dataset D based on the boundary knowledge database K_B. Specifically, to build the boundary knowledge database K_B, we leveraged 4,814 authoritative encyclopedic entries in the LIS domain and the bibliographic data of 93,971 Chinese Social Sciences Citation Index (CSSCI)-indexed journal articles published between 1998 and 2023. Through boundary knowledge enhancement and validation using multiple top-tier large language models, we constructed 276,083 question-answer pairs as the training set for LISGPT. Subsequently, based on WSRS sampling, we generated 1,001 positive and negative example data samples for reinforcement learning from human feedback by maliciously rewriting using the GPT-3.5-turbo model.

To evaluate the training effectiveness of LISGPT, we conducted tests on two tasks: keyword prediction and professional translation. For this purpose, we constructed corresponding test sets using the bibliographic data of 1,447 CSSCI-indexed journal articles published in the LIS domain in 2024.

Model Training Parameters

LISGPT was trained on three different specifications of the Qwen instruct-tuned models: 3B, 7B, and 14B. First, we initialized the models using the MS-SWIFT framework (Zhao et al., 2024) to perform supervised fine-tuning on the Qwen2.5 models based on the LoRA algorithm. The data for LoRA fine-tuning was derived from the boundary knowledge-enhanced question-answer dataset D, aiming to expand the model's breadth and depth of knowledge in the LIS domain. The training configuration included 1 epoch (num_train_epochs), a batch size of 1, and a learning rate of 5e-5. Next, we conducted model optimization based on DBKO. This phase utilized two key datasets: the question-answer data A_p and $\hat{A}_s$, which were extracted and rewritten using the WSRS method. Training was performed using bfloat16 mixed precision, with the adamw_torch optimizer and a cosine learning rate scheduler. The learning rate was set to 5e-5, with a gradient accumulation step of 8 and a per-device batch size of 1. The number of training epochs was set to 3.0. For the DPO loss function the specific hyperparameters were configured as follows: pref_beta was set to 0.1, pref_ftx to 0.1, and the preference loss used the sigmoid function. The entire training process was executed on a Linux server equipped with 4 NVIDIA RTX 4090 (24GB) GPUs.

Empirical Results

Keyword Prediction

This study evaluated LISGPT for keyword prediction against SOTA baseline models, and the results demonstrate its significant effectiveness in predicting keywords based on paper titles. As shown in Table 1, among all evaluated models, LISGPT-14b achieved the highest F1 score (0.3973), surpassing even large commercial models such as DeepSeek-V3 and Qwen-Max. The outstanding performance of LISGPT-14b is primarily attributed to its exceptional precision (0.4247), which ranks first among all models, indicating its ability to generate highly accurate and relevant outputs. Notably, even the smaller variants of LISGPT in this study, LISGPT-7b and LISGPT-3b, demonstrated remarkable efficiency. They ranked second (0.4219) and third (0.4216) in precision, respectively, and third (0.3827) and fourth (0.3789) in overall F1 scores, outperforming larger models such as Qwen2.5-Max (0.3712), Claude-3-Haiku (0.3571), and GPT-4o-mini (0.3423). This indicates that the architecture and training methods proposed in this study can effectively capture the subtleties of title comprehension and synthesis without necessarily requiring a large number of parameters.

Although DeepSeek-V3 achieved the highest recall (0.3979), with LISGPT-14b ranking fourth in this metric (0.3733), the balance between precision and recall in the F1 score favors LISGPT-14b. This suggests that while responses generated by LISGPT-14b may occasionally omit certain elements of an instruction, they contain fewer errors or irrelevant components compared to other models. Collectively, these results indicate that LISGPT demonstrates a strong understanding of LIS domain knowledge and outperforms current SOTA models across various parameter ranges.

Model	Precision	Rank(P)	Recall	Rank(R)	F1	Rank(F1)
LISGPT-14b	0.4247	1	0.3733	4	0.3973	**1**
DeepSeek-V3	0.3784	6	0.3979	1	0.3879	2
LISGPT-7b	0.4219	**2**	0.3502	**7**	0.3827	**3**
LISGPT-3b	0.4216	**3**	0.344	**9**	0.3789	**4**
Qwen2.5-7b-Instruct	0.4022	5	0.355	6	0.3771	5
Qwen2.5-max	0.3528	9	0.3916	2	0.3712	8
Doubao-1.5-Lite	0.4027	4	0.3449	8	0.3716	7
Qwen2.5-14b-Instruct	0.3666	8	0.383	3	0.3746	6
Claude-3-Haiku	0.3779	7	0.3384	10	0.3571	9
GPT-4o mini	0.3164	11	0.3728	5	0.3423	10
Qwen2.5-3b-Instruct	0.3267	10	0.2983	11	0.3119	11
Llama-3.1-8B-Instruct	0.1663	12	0.1856	12	0.1754	12
ERNIE-Tiny-8K	0.118	13	0.118	13	0.118	13

Table 1. Performance comparison between LISGPT and general LLMs in keyword prediction.

Professional Translation

In the domain of professional translation, we conducted a comprehensive evaluation of LISGPT against baseline models, yielding highly encouraging results. As shown in Table 2, LISGPT-14b achieved the third-highest overall performance (average = 0.5971) among all evaluated models, reaching 99.1% of the performance level of DeepSeek-V3 (0.6025). This indicates that LISGPT-14b is highly competitive in this specialized translation task. Although DeepSeek-V3 and Qwen2.5-Max slightly outperformed LISGPT-14b, the performance gap was notably small. LISGPT-14b demonstrated strong BLEU performance (0.3681), ranking second among all models, just slightly behind Qwen2.5-Max (0.3685). Similarly, it achieved the second-highest ROUGE-1 score (0.7778), indicating its ability to maintain high lexical overlap with reference translations. Particularly noteworthy are the performances of LISGPT-7b and LISGPT-3b, which ranked fifth (0.5797) and sixth (0.5732) in overall scores, respectively. These smaller variants outperformed larger commercial models, including GPT-4o-mini (0.5728), Claude-3-Haiku (0.5517), and all variants of the Qwen2.5 series. The robust performance of the smaller model variants highlights the efficiency of our approach, demonstrating that LISGPT's architecture can effectively capture the subtleties of professional terminology translation without requiring excessive parameters. Further validation of LISGPT's effectiveness is provided by its ROUGE scores. LISGPT-14b ranked second, fourth, and fourth in the ROUGE-1, ROUGE-2, and ROUGE-L metrics, respectively, showing consistent performance across different evaluation dimensions. This indicates that the translations generated by LISGPT-14b not only cover reference terms well but also preserve appropriate sequential structures.

Collectively, these empirical results demonstrate that LISGPT models can effectively address the challenges of professional terminology translation, exhibiting highly competitive performance compared to leading commercial models. They also indicate that the model architecture and training methods developed in this study achieve translation quality comparable to larger commercial models while maintaining a smaller model size, enabling cross-lingual knowledge connectivity. While there is still room for improvement in capturing fine syntactic structures of professional terminology, LISGPT represents a significant advancement in professional language translation capabilities.

Model	Avg_BLEU	Avg_ROUGE1	Avg_ROUGE2	Avg_ROUGEL	Average	Rank
DeepSeek-V3	0.3674	0.7789	0.5593	0.7042	0.6025	1
Qwen2.5-max	0.3685	0.776	0.5593	0.7019	0.6014	2
LISGPT-14b	0.3681	0.7778	0.5477	0.6948	0.5971	**3**
Doubao-1.5-Lite	0.352	0.7726	0.552	0.7011	0.5944	4
LISGPT-7b	0.343	0.7664	0.5264	0.6829	0.5797	**5**
LISGPT-3b	0.3409	0.7577	0.5205	0.6735	0.5732	**6**
GPT-4o mini	0.3303	0.7568	0.5235	0.6805	0.5728	7
Claude-3-Haiku	0.3058	0.7416	0.4989	0.6603	0.5517	8
Qwen2.5-7b-Instruct	0.2902	0.7253	0.4743	0.637	0.5317	9
Qwen2.5-14b-Instruct	0.2697	0.7111	0.457	0.6278	0.5164	10
ERNIE-Tiny-8K	0.208	0.6849	0.439	0.5924	0.4811	11
Qwen2.5-3b-Instruct	0.2376	0.685	0.4154	0.5862	0.4811	12
Llama-3.1-8B-Instruct	0.2098	0.6716	0.392	0.5462	0.4549	13

Table 2. Performance comparison between LISGPT and general LLMs in professional translation.

Ablation Study

To validate the effectiveness of the DBKO strategy, we compared the performance of LISGPT models with and without this strategy. Table 3 presents the detailed comparison results.

Model	Keyword Prediction			Professional Translation				
	Precision	Recall	F1	Avg_BLEU	Avg_ROUGE1	Avg_ROUGE2	Avg_ROUGEL	Average
LISGPT-14b-without_DBKO	0.4222	0.3808	0.4004	0.3422	0.7544	0.5282	0.6768	0.5754
LISGPT-7b-without_DBKO	0.4031	0.3715	0.3867	0.2912	0.7664	0.5259	0.6841	0.5669
LISGPT-3b-without_DBKO	0.3912	0.3700	0.3803	0.2878	0.7576	0.5192	0.6762	0.5602
LISGPT-14b	0.4247	0.3733	0.3973	0.3681	0.7778	0.5477	0.6948	0.5971
LISGPT-7b	0.4219	0.3502	0.3827	0.3430	0.7664	0.5264	0.6829	0.5797
LISGPT-3b	0.4216	0.3440	0.3789	0.3409	0.7577	0.5205	0.6735	0.5732

Table 3. Ablation study result.

In the keyword prediction task, the LISGPT-14b model with DBKO showed an improvement in the precision metric by 0.0025 (from 0.4222 to 0.4247) compared to the version without DBKO. However, there was a slight decrease in the recall metric (from 0.3808 to 0.3733), resulting in a marginal reduction in the overall F1 score (from 0.4004 to 0.3973). Similarly, for LISGPT-7b and LISGPT-3b, the application of DBKO led to significant improvements in precision, albeit with small decreases in recall and F1 score. These results indicate that the DBKO strategy effectively enhances the accuracy of the model's generated content but may slightly reduce its coverage scope.

The impact of DBKO was even more pronounced in the professional title translation task. After applying DBKO, the LISGPT-14b model achieved notable improvements across all metrics: BLEU score increased by 0.0259 (from 0.3422 to 0.3681), ROUGE-1 improved by 0.0234 (from 0.7544 to 0.7778), ROUGE-2 increased by 0.0195 (from 0.5282 to 0.5477), ROUGE-L improved by 0.0180 (from 0.6768 to 0.6948).

The overall average score increased by 0.0217 (from 0.5754 to 0.5971), representing a 3.77% improvement. For LISGPT-7b and LISGPT-3b, consistent performance gains were observed as well. Their BLEU scores improved by more than 0.05, and their overall average scores increased by over 0.01, corresponding to an improvement of more than 2%. These results clearly demonstrate the effectiveness of the DBKO strategy in knowledge-intensive tasks, particularly in enhancing the precision and professionalism of generated content.

CONCLUSION

This study introduces LISGPT and openly provides its dataset and models to promote future research in the fields.

We designed the **B**oundary **K**nowledge **E**nhance (BKE) framework, which incorporates authoritative LIS knowledge and high-quality journal papers as a knowledge database. Combined with question generation driven by Bloom's Taxonomy and retrieval-augmented response mechanisms, we constructed a high-quality, large-scale LIS professional question-answer dataset. Based on this dataset, we proposed the **D**irect **B**oundary **K**nowledge **O**ptimization (DBKO) training method, which uses the **W**eighted **S**quare **R**atio **S**ampling (WSRS) technique to further construct preference training samples from high-quality question-answer pairs. The LISGPT series of models were trained on different-sized base models (Qwen2.5-3b/7b/14b-Instruct). LISGPT demonstrated superior performance compared to SOTA commercial models. It outperformed all baseline models in the literature keyword prediction task (F1 = 0.3973, rank = 1) and achieved 99.1% of the performance level of DeepSeek-V3-671b in the professional translation task (average = 0.5971, rank = 3). Notably, even smaller-parameter variants like LISGPT-7b and LISGPT-3b achieved remarkable results. Additionally, under local server execution conditions, LISGPT demonstrated significant economic advantages with low training, deployment, and inference costs, substantially reducing the application cost of domain-specific models. This provides affordable intelligent support tools for both upstream and downstream tasks such as LIS knowledge extraction and academic evaluation.

The significance of this study lies in providing a feasible paradigm for constructing specialized large models in the social sciences. The BKE framework and DBKO method can create high-quality and diverse datasets for social science reasoning tasks, addressing challenges such as limited generation diversity and difficulty in validating generated data. These approaches are not only applicable to LIS but can also be extended to other humanities and social sciences, providing important insights for building open-source, transparent, and efficient domain-specific large language models.

GENERATIVE AI USE

We employed Qwen2.5-Max for the following purposes: (1) translating parts of sections of the text into English, (2) proofreading and correcting grammatical errors, and (3) refining the phrasing and style of the manuscript to enhance clarity and readability. We evaluated the output by cross-referencing the translated and revised content with the original text to ensure accuracy, consistency, and alignment with the intended meaning. Additionally, we reviewed the final version to confirm that all technical terms and concepts were appropriately conveyed. The authors assume all responsibility for the content of this submission.

AUTHOR ATTRIBUTION

First Author: conceptualization, methodology, visualization, formal analysis, writing – original draft, writing – review and editing; Second Author: methodology, data curation, writing – original draft, writing – review and editing; Third Author: methodology, writing – original draft; Fourth Author: supervision; Fifth Author: supervision, funding acquisition, project administration.

ACKNOWLEDGMENTS

Authors gratefully acknowledge the grant from the major project of the National Social Science Foundation of China (Grant No. 24&ZD323) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX25_0130).

REFERENCES

Bloom, B. S. (1984). *Taxonomy of educational objectives, handbook 1: Cognitive domain* (2nd edition Edition). Addison-Wesley Longman Ltd.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *137*, 85–86.

CharityGPT. (2024). *Scholar GPT*. Scholar GPT. https://www.charitygpt.io/gpts/scholar-gpt

Chen, J., Liu, L., & Zhou, F. (2025). Do not wait: Preemptive rumor detection with cooperative LLMs and accessible social context. *Information Processing & Management*, *62*(3), 103995. https://doi.org/10.1016/j.ipm.2024.103995

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024, June 6). *Chatbot arena: An open platform for evaluating LLMs by human preference*. Forty-first International Conference on Machine Learning. https://openreview.net/forum?id=3MW8GKNyzI

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53.

Clarivate. (2024, September 4). *Clarivate launches generative AI-powered web of science research assistant | clarivate*. https://clarivate.com/news/clarivate-launches-generative-ai-powered-web-of-science-research-assistant/

CNKI. (2024, April 28). *The Chinese Knowledge Large Language Model*. Huazhi Wensi. https://huazhi.cnki.net

Cuskley, C., Woods, R., & Flaherty, M. (2024). The limitations of large language models for understanding human language and cognition. *Open Mind*, *8*, 1058–1083. https://doi.org/10.1162/opmi_a_00160

Dervin, B. (1998). Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, *2*(2), 36–46. https://doi.org/10.1108/13673279810249369

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*. https://doi.org/10.1126/science.adi1778

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, October 6). *LoRA: Low-Rank Adaptation of Large Language Models*. International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9

Lee, J., Stevens, N., & Han, S. C. (2025). Large language models in finance (FinLLMs). *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10495-6

Li, B., Fan, S., & Huang, J. (2024). CSAFT: Continuous semantic augmentation fine-tuning for legal large language models. In M. Wand, K. Malinovská, J. Schmidhuber, & I. V. Tetko (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2024* (pp. 293–307). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72344-5_20

Li, M., Zhao, Y., Deng, Y., Zhang, W., Li, S., Xie, W., Ng, S.-K., & Chua, T.-S. (2024). *Knowledge boundary of large language models: A survey* (No. arXiv:2412.12472). arXiv. https://doi.org/10.48550/arXiv.2412.12472

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., … Zhao, L. (2024). *Domain specialization as the key to make large language models disruptive: A comprehensive survey* (No. arXiv:2305.18703). arXiv. https://doi.org/10.48550/arXiv.2305.18703

Liu, Y., Bu, N., Li, Z., Zhang, Y., & Zhao, Z. (2025). AT-FinGPT: Financial risk prediction via an audio-text large language model. *Finance Research Letters*, *77*, 106967. https://doi.org/10.1016/j.frl.2025.106967

Lu, H., Liu, T., Cong, R., Yang, J., Gan, Q., Fang, W., & Wu, X. (2025). QAIE: LLM-based quantity augmentation and information enhancement for few-shot aspect-based sentiment analysis. *Information Processing & Management*, *62*(1), 103917. https://doi.org/10.1016/j.ipm.2024.103917

Qian, L., Zhang, Z., Wu, D., Chang, Z., Yu, Q., Hu, M., & Liu, Y. (2024). The Large Language Model for Scientific Literature: Method, Framework, and Application. *Journal of Library Science in China*, *50*(6), 45–58. https://doi.org/10.13530/j.cnki.jlis.2024046

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*, 53728–53741.

Sadia, B., Adeeba, F., Shams, S., & Hussain, S. (2025). Leveraging LLMs for action item identification in Urdu meetings: Dataset creation and comparative analysis. *Information Processing & Management*, *62*(3), 104071. https://doi.org/10.1016/j.ipm.2025.104071

Shi, J., Guo, Q., Liao, Y., Wang, Y., Chen, S., & Liang, S. (2024). Legal-LM: Knowledge graph enhanced large language models for law consulting. In D.-S. Huang, Z. Si, & C. Zhang (Eds.), *Advanced Intelligent Computing Technology and Applications* (pp. 175–186). Springer Nature. https://doi.org/10.1007/978-981-97-5672-8_15

Suzuki, M., Sakaji, H., Hirano, M., & Izumi, K. (2023). Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management*, *60*(2), 103194. https://doi.org/10.1016/j.ipm.2022.103194

Thelwall, M. (2025). ChatGPT for complex text evaluation tasks. *Journal of the Association for Information Science and Technology*, *76*(4), 645–648. https://doi.org/10.1002/asi.24966

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Wang, N., Yang, H., & Wang, C. (2023, November 26). *FinGPT: Instruction tuning benchmark for open-source large language models in financial datasets*. NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following. https://openreview.net/forum?id=FuOMomaQa8

Whitfield, S., & Hofmann, M. A. (2023). Elicit: AI literature review research assistant. *Public Services Quarterly*. https://www.tandfonline.com/doi/abs/10.1080/15228959.2023.2224125

Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., & Wang, Y. (2024). PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, *31*(9), 1833–1843. https://doi.org/10.1093/jamia/ocae045

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, *36*, 33469–33484.

Yao, S., Ke, Q., Wang, Q., Li, K., & Hu, J. (2024). Lawyer GPT: A legal large language model with enhanced domain knowledge and reasoning capabilities. *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, 108–112. https://doi.org/10.1145/3689299.3689319

Ye, G., Zhao, H., Zhang, Z., & Jiang, Z. (2025). UniDE: A multi-level and low-resource framework for automatic dialogue evaluation via LLM-based data augmentation and multitask learning. *Information Processing & Management*, *62*(3), 104035. https://doi.org/10.1016/j.ipm.2024.104035

Yin, X., Zhang, X., Ruan, J., & Wan, X. (2024). Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers)* (pp. 2270–2286). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.124

Yu, F., Kincaide, H., & Carlson, R. B. (2024). An empirical study evaluating ChatGPT's performance in generating search strategies for systematic reviews. *Proceedings of the Association for Information Science and Technology*, *61*(1), 423–434. https://doi.org/10.1002/pra2.1039

Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., Zhou, W., & Chen, Y. (2024). *SWIFT:a scalable lightWeight infrastructure for fine-tuning* (No. arXiv:2408.05517). arXiv. https://doi.org/10.48550/arXiv.2408.05517

Zhou, Z., Yu, K.-Y., Tian, S.-Y., Yang, X.-W., Shi, J.-X., Song, P., Jin, Y.-X., Guo, L.-Z., & Li, Y.-F. (2025). *LawGPT: Knowledge-guided data generation and its application to legal LLM* (No. arXiv:2502.06572). arXiv. https://doi.org/10.48550/arXiv.2502.06572

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1). https://doi.org/10.1162/coli_a_00502