• RESEARCH PAPER •

Knowledge-Driven Agentic Scientific Corpus Distillation Framework for Biomedical Large Language Models Training

Meng Xiao[1,3], Xunxin Cai[1,2], Qingqing Long[1], Chengrui Wang[1], Yuanchun Zhou[1,2]& Hengshu Zhu[1*]

[1]*Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*
[2]*University of Chinese Academy of Sciences, Beijing, China*
[3]*Duke-NUS Medical School, National University of Singapore, Singapore*

Abstract Corpus distillation for biomedical large language models (LLMs) seeks to address the pressing challenge of insufficient quantity and quality in open-source annotated scientific corpora, which remains a bottleneck for effective LLM training in biomedical research. This paper proposes a knowledge-driven, agentic framework for scientific corpus distillation, tailored explicitly for LLM training in the biomedical domain, addressing the challenge posed by the complex hierarchy of biomedical knowledge. Central to our approach is a collaborative multi-agent architecture, where specialized agents—each guided by the Medical Subject Headings (MeSH) hierarchy—work in concert to autonomously extract, synthesize, and self-evaluate high-quality textual data from vast scientific literature. This agentic framework collectively generates and refines domain-specific question-answer pairs, ensuring comprehensive coverage and consistency with biomedical ontologies while minimizing manual involvement. Extensive experimental results show that language models trained on our multi-agent distilled datasets achieve notable improvements in biomedical question-answering tasks, outperforming both strong life sciences LLM baselines and advanced proprietary models. Notably, our AI-Ready dataset enables Llama3-70B to surpass GPT-4 with MedPrompt and Med-PaLM-2, despite their larger scale. Detailed ablation studies and case analyses further validate the effectiveness and synergy of each agent within the framework, highlighting the potential of multi-agent collaboration in biomedical LLM training.*

Keywords Biomedical Large Language Models, Agentic Corpus Distillation, Synthetic Question–Answer Generation, Agentic AI, Knowledge Hierarchy Guidance

Citation Knowledge-Driven Agentic Scientific Corpus Distillation Framework for Biomedical Large Language Models Training. Sci China Inf Sci, for review

1 Introduction

The advent of large language models has propelled bioinformatics into a new era [1], enabling the development of automated solutions across a spectrum of biomedical domains [2–5], and has demonstrated notable success for real-world biomedical question answering (QA) tasks [6–8]. However, the intricate and specialized nature of biomedical tasks means that general-purpose LLMs often fall short unless meticulously adapted and fine-tuned for the domain [9–11]. Progress in this area is further constrained by the scarcity of sufficiently large and high-quality biomedical corpora [12–14]. While existing open-source biomedical datasets are typically of high quality, their limited scale and narrow topical coverage restrict their utility for comprehensive LLM training. Conversely, directly leveraging the expansive body of domain-specific scientific literature offers the potential to cover a broader range of biomedical topics. Nevertheless, most of these resources are unannotated, and their inherent lack of structure, coupled with the absence of QA-format organization, greatly hinders their effective use for training question-answering models. Although the vast body of biomedical literature represents a valuable and authoritative resource, its complex terminology and dense conceptual structures pose significant barriers to automated processing and dataset construction [15, 16]. Figure 1 visualizes the resulting data bottleneck. These challenges raise a crucial question: **How can we automatically distill large-scale, high-quality QA pairs from scientific literature to empower biomedical LLM?**
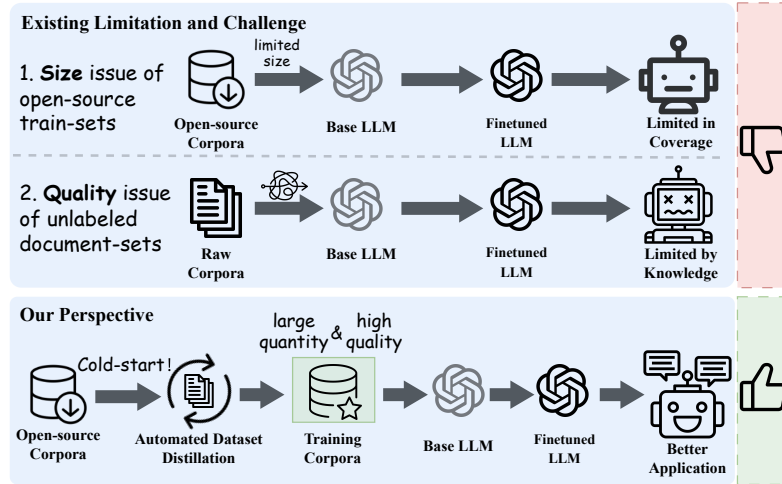
Figure 1 Analyzing the limitations and challenges of the existing pipeline. The motivation of this study is to utilize the high-quality but limited annotated corpus to generate large-scale training corpora from raw scientific documents.

Prior research can be categorized into three mainstream categories. First, rule-based methods [17–19] rely on human-crafted standards for data cleaning and curation, which, while effective in reducing noise, are resource-intensive and difficult to scale. Second, knowledge graph-based approaches [20, 21] structure biomedical information from texts into comprehensive graphs, but their dependence on manually curated sources limits both efficiency and scalability. Third, synthetic data generation methods [22–26] use LLMs to automate the creation of QA pairs and process large corpora. Yet, they often lack mechanisms for interdisciplinary collaboration [27, 28], resulting in insufficient diversity and robustness in the distilled data. Recent efforts [29] have introduced a knowledge hierarchy-guided [30] approach that leverages a single LLM agent to generate and evaluate biomedical QA data with improved alignment to domain ontologies. Its reliance on a single-agent, rule-based architecture presents inherent limitations in terms of diversity, cross-domain expertise, and collaborative reasoning [31].

In response to these limitations, we introduce the **M**ulti-agent enhanced **K**nowledge hier**A**rchy gu**I**ded biomedical dataset disti**L**lat**I**o**N** (**m-KAILIN**) framework—a novel, fully automated, agentic framework designed to extract high-quality training corpora for biomedical LLMs. The m-KAILIN framework leverages a collaborative architecture of specialized LLM agents, each guided by a structured biomedical knowledge hierarchy, to evaluate and generate 'Question-Answer-Context' triples with strong domain alignment. The workflow begins with fine-tuned LLM agents generating candidate questions from limited annotated datasets. These questions are then matched to relevant contexts retrieved from over 23 million biomedical research articles. The question-context pair that best aligns with the knowledge hierarchy is chosen through a multi-agent evaluation and selection process. This automated pipeline facilitates the generation of high-quality, preference-based datasets that can further train LLMs to produce questions and answers aligned with biomedical ontologies, culminating in a robust, AI-ready corpus.

In summary, our contributions are listed as follows:

• **Automated Biomedical Corpus Distillation Workflow**: We present a multi-agent workflow that enables the fully automatic extraction and distillation of biomedical corpora from large-scale research literature, vastly reducing the need for manual annotation while improving efficiency and dataset coverage.

• **Framework and Methodology**: Our framework introduces a knowledge hierarchy-driven evaluation mechanism, leveraging MeSH to guide and assess the multi-agent distillation. This ensures the extracted data is both contextually relevant and domain-consistent, eliminating the need for human curation.

• **Comprehensive Empirical Validation**: We conduct extensive experiments to validate the effectiveness of m-KAILIN and the quality of the generated datasets. Detailed ablation studies and case analyses reveal the contributions of each framework component, and our investigation of data scaling laws provides actionable insights into corpus distillation strategies for biomedical LLMs.

2 Related Work

2.1 Dataset Distillation.

Dataset distillation [32, 33] is an information extraction technique that uses generative models to distill core information from large-scale raw datasets, resulting in high-quality datasets. Existing dataset distillation approaches are primarily categorized based on their optimization objectives into three types: performance matching, parameter matching, and distribution matching. Performance matching methods [34–37] aim to optimize dataset distillation by minimizing the loss of a model trained on the distilled dataset when evaluated on the raw dataset. Parameter matching methods [38–41] focus on the consistency of trainable model parameters when trained on the distilled dataset compared to the raw dataset. Unlike these two training-based objectives, distribution matching methods [42] analyze the distilled dataset itself, using the distributional consistency between the distilled and raw datasets as the target for optimization.

2.2 Synthesis Text Data in Biomedical field.

Existing methods for generating and utilizing synthetic text data in biomedical text mining face several challenges. PubMedQA [43] relies on converting titles into questions, often introducing noise due to simplistic or irrelevant content. Attempts to improve PubMedQA through LLM-based rewriting strategies, like those using GPT-3.5-turbo or GPT-4, fail to offer significant diversity in generated content [44]. MedSyn [22], while leveraging a Medical Knowledge Graph and LLMs like GPT-4, lacks a comprehensive integration of hierarchical knowledge. Some other methods that rely on experts to use LLMs with their own knowledge [23, 45] place too much dependence on expert effort, lacking automation and efficiency. Besides, KAILIN [29] first introduced the MeSH knowledge hierarchy as a rule-based evaluator, enhancing the quality of the generated questions. In this study, we advance beyond KAILIN by proposing an agentic, collaborative framework, where multiple specialized agents—each guided by biomedical ontologies—autonomously generate, evaluate, and refine the generated dataset.

2.3 LLMs in Biomedical Domain.

Recent advancements [46, 47] in biomedical language models have focused on improving domain adaptation to better handle dense terminology and complex medical concepts. Approaches such as retrieval-augmented generation have enhanced models' ability to retrieve relevant information for domain-specific tasks [48–50]. PMC-LLaMA [20] has been fine-tuned using millions of biomedical papers and medical textbooks, integrating medical knowledge graphs like UMLS for domain knowledge. Similarly, BioMistral [17] has applied extensive data processing and adaptive pretraining on PubMed Central, while HEAL [51] has combined public and proprietary datasets for continuous pretraining on a general-purpose model, further enhancing its biomedical capabilities.

3 Methods

In this section, we detail the m-KAILIN framework, which leverages an agentic framework design to distill biomedical corpora from large-scale literature systematically.

3.1 Question Generation Agent

The Question Generation Agent is responsible for transforming biomedical documents into high-quality, domain-specific questions. This agent operates in two stages: (1) fine-tuning large language models for biomedical question generation, and (2) utilizing the trained models to generate candidate questions from unseen documents.

3.1.1 *Fine-Tuning on Open-Source QA Dataset*

To enable effective biomedical question generation, we fine-tune a large language model on the BioASQ QA dataset [52], and resulting in a specialized generator, θ. Let $\mathcal{T} = \{(d_i, q_i)\}_{i=1}^{N}$ denote the open-source

QA training set, where d_i is a document and q_i its reference question. The objective function is:

$$\mathcal{L}_{\text{QA}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log P_\theta(q_i|d_i), \tag{1}$$

where $P_\theta(q_i|d_i)$ is the model's predicted probability of the target question q_i. This cross-entropy loss ensures θ learn to generate questions aligned with biomedical knowledge.

3.1.2 *Question Generation from General Biomedical Documents*

During inference, each trained question generator receives a biomedical document d_i sampled from PubMed and generates a corresponding candidate question q_i. Formally,

$$q_i = \arg\max_q P_\theta(q_i|d_i) \tag{2}$$

where θ denotes the fine-tuned question generator. This process enables the framework to automatically produce diverse and domain-relevant questions from large-scale biomedical literature, laying the groundwork for subsequent context retrieval and evaluation stages.

3.2 Context Retrieval Agent

The Context Retrieval Agent retrieves relevant supporting documents from the large-scale PubMed[1] corpus for each generated question. This agent employs Dense Passage Retrieval (DPR) [53], in line with Retrieval-Augmented Generation (RAG) [54] frameworks. For each question q_i, we encode both the question and every candidate document c as dense vectors, $\mathbf{e}(q_i)$ and $\mathbf{e}(c)$, using a shared BiomedBERT$_{\text{base}}$ encoder[2]. The relevance score is computed as:

$$p(c|q_i) \propto \exp\left(\mathbf{e}(c)^\top \mathbf{e}(q_i)\right), \tag{3}$$

Documents with top-k relevance scores are selected as contexts $\mathbf{c}_i = \{c_1, c_2, ..., c_k\}$ for subsequent evaluation. This decoupled retrieval enables independent optimization and clearer agent collaboration in the overall framework.

3.3 Question Evaluation Agent

The Question Evaluation Agent is designed to assess which of two candidate question-context pairs, $(q_i^a, \mathbf{c}_i^a)$ and $(q_i^b, \mathbf{c}_i^b)$, generated from the same source document d_i, better aligns with the hierarchical biomedical knowledge embodied in Medical Subject Headings [55] (MeSH). The agent employs a two-stage approach: an initial rule-based evaluation leveraging the MeSH ontology to create high-quality preference labels, followed by learning an automatic evaluator via LLM fine-tuning.

3.3.1 *Knowledge-Driven Fine-Tuning*

To enable evaluation without relying on human-annotated preference data, we propose a knowledge-guided, rule-based cold-start method to compare candidate question-context pairs. This approach leverages the structural and semantic information from the MeSH ontology to generate high-confidence preference labels at scale. Given a source document d_i and two candidate pairs $(q_i^a, \mathbf{c}_i^a)$ and $(q_i^b, \mathbf{c}_i^b)$, we can obtain MeSH terms[3] from d_i, $\mathbf{c}_i^a$, and $\mathbf{c}_i^b$. Referring to Lin's approach [56], for given context set $\mathbf{c}_i$, the knowledge hierarchy similarity to d_i is computed as:

$$\bar{S}_i^j = \frac{1}{|d_i| |\mathbf{c}_i|} \sum_{m^x \in d_i} \sum_{m^y \in \mathbf{c}_i} \frac{2 \times IC(\Lambda(m^x, m^y))}{IC(m^x) + IC(m^y)}, \tag{4}$$

---

where $IC(m)$ is the information content for MeSH term m, and $\Lambda(m^x, m^y)$ is their lowest common ancestor. For any structured subject term m, we first calculate its information content as:

$$IC(m) = -\log(\frac{freq(\mathcal{M}(m))}{n_{terms}}), \tag{5}$$

where $\mathcal{M}(m)$ denotes the set of all descendants of MeSH term m, and n_{terms} represents the total number of MeSH terms in the corpus. This score reflects how well the candidate context's MeSH terms align with those of the source document. A higher $\bar{S}_i^j$ indicates that the candidate context is more semantically consistent with the document's knowledge hierarchy. Once the similarity scores for each candidate pair have been calculated, we select the one with the higher score:

$$y_i = \arg\max_{j \in \{a,b\}} \bar{S}_i^j \tag{6}$$

In other words, the candidate pair whose context set exhibits a stronger semantic match to the source document is preferred. These cold-start preference labels, obtained in a rule-based and knowledge-guided manner, provide a large-scale and high-quality supervision signal. This signal can then be used to bootstrap a subsequent LLM-based evaluation agent.

3.3.2 *Evaluation Agent Preference Learning*

With the knowledge-guided cold-start preference labels as supervision, we further train a large language model ϕ to serve as an automatic evaluation agent. Given a training set of triplets $\{(d_i, (q_i^a, \mathbf{c}_i^a), (q_i^b, \mathbf{c}_i^b), y_i)\}$, where y_i indicates which candidate pair is preferred according to the MeSH-based scoring, the LLM is optimized to predict the correct preference. Formally, let $P_\phi(y_i | d_i, q_i^a, \mathbf{c}_i^a, q_i^b, \mathbf{c}_i^b)$ denote the predicted probability that the y_i-th pair is preferred. The LLM is trained by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{eval}}(\phi) = -\frac{1}{N} \sum_{i=1}^{N} \log P_\phi(y_i | d_i, q_i^a, \mathbf{c}_i^a, q_i^b, \mathbf{c}_i^b), \tag{7}$$

where this process enables the Question Evaluation Agent to learn preference judgment criteria derived from domain knowledge, without requiring human-annotated data.

### 3.3.3 *Automatic Evaluation and Deployment*

Once fine-tuned, the LLM-based evaluation agent can be deployed to automatically assess the relative quality of any two candidate question-context pairs generated from the same document. Given $(d_i, (q_i^a, c_i^a), (q_i^b, c_i^b))$, the agent predicts which pair is better aligned with biomedical knowledge, thus supporting scalable and efficient data distillation for downstream tasks. This learned evaluation mechanism inherits the domain expertise embedded in the MeSH-based rules and generalizes to more complex or subtle cases, enabling robust multi-agent collaboration in the m-KAILIN framework.

3.4 Answer Generation Agent

Once the optimal question-context pair is identified, the Answer Generation Agent φ generates high-quality answers using advanced LLMs such as GPT-4o [57]. For each input $(q_j, \mathbf{c}_j)$, this agent produces an answer $\varphi(q_j, \mathbf{c}_j) \to a_j$, facilitating the construction of comprehensive QA datasets for downstream training.

3.5 Agentic Collaborative Framework

This stage integrates outputs from the multiple agents to construct datasets, and further optimizes the question generator through direct preference optimization [58] (DPO):
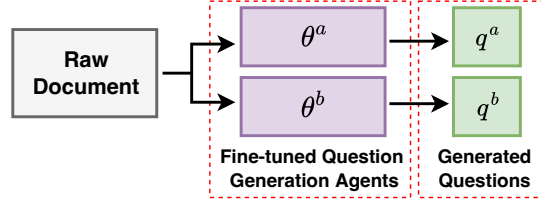
Figure 2 The two distinct Question Generation Agents will generate different question by given the raw documents.

3.5.1 *Initialize Two Distinct Question Generation Agents*

As depicted in Figure 2, m-KAILIN framework will first initialize two Question Generation Agents. Specifically, one of them, θ^a, will be based on a domain-specific LLM, e.g., BioMistral [17]. And the other θ^b will be based on a more general yet powerful LLM, e.g., Llama-3 [59]. Each of them will follow the same fine-tuning step introduced in Section 3.1.1 and be able to generate quality questions. We want to highlight that the reason we adopt two distinct Question Generation Agents is to enhance diversity and robustness in question generation. The domain-specific agent (e.g., BioMistral) is expected to capture fine-grained biomedical knowledge and terminology, ensuring that generated questions are highly relevant to the biomedical context. Meanwhile, the general-purpose agent (e.g., Llama-3) possesses broader reasoning and linguistic capabilities and may generate questions from more diverse perspectives. By leveraging both agents, m-KAILIN can comprehensively explore the question space over biomedical documents, increasing the likelihood of producing high-quality, informative, and non-redundant question candidates.
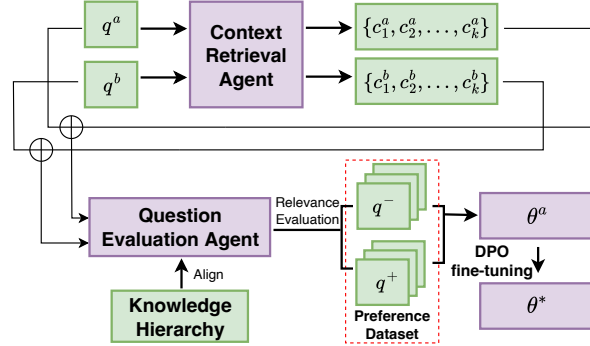
Figure 3 The generated question pairs will be evaluated based on how well their retrieved contexts align with the knowledge hierarchy of the raw document.

3.5.2 *Retrieve Context of Generated Question*

The outputs from both agents will be used in the subsequent context retrieval and evaluation steps, allowing the downstream modules to select the most suitable question-context pairs for the construction of high-quality training datasets. Specifically, each pair of generated questions will feed into the Context Retrieval Agent to obtain their top-k relevant documents as supplementary context (as depicted in the top region of Figure 3).

### 3.5.3 *Preference Dataset Construction*

By utilizing the Evaluation Agent for each generated pair from two Question Generation Agents, we select the question-context pair with the better alignment as the positive sample (q_i^+), and the other as the negative (q_i^-), forming a preference dataset $\mathcal{P}$ (highlighted by the red dashed line in Figure 3). This preference dataset forms the supervision signal for optimizing the question generation agent in the next step.

3.5.4 *Direct Preference Optimization*

The Direct Preference Optimization (DPO) objective explicitly aligns the model's output distribution with automated preference signals. Given a set of preference data $\{(d_i, q_i^+, q_i^-)\}$, where q_i^+ is the preferred question over q_i^- for document d_i, DPO fine-tunes the general LLM θ^b to maximize the likelihood of generating preferred questions while minimizing the likelihood of dispreferred ones. Formally, the DPO objective is:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\beta\log P_\theta(q_i^+|d_i)\right)}{\exp\left(\beta\log P_\theta(q_i^+|d_i)\right) + \exp\left(\beta\log P_\theta(q_i^-|d_i)\right)} \tag{8}$$

where β is a temperature scaling factor. The optimized model θ^* is thus directly steered towards generating questions that align with preference signals and can serve as the basis for constructing high-quality synthetic datasets in the final step.
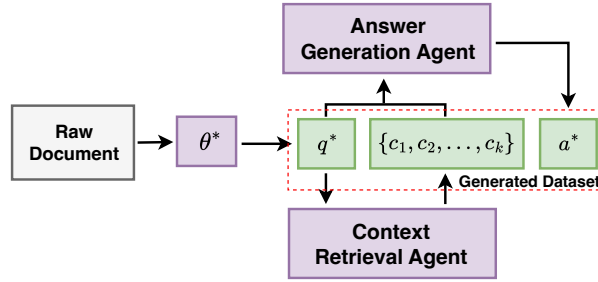
Figure 4 The fine-tuned Question Generation Agent will collaorated with Answer Generation Agent and Context Retrieval Agent to build the training corpora dataset.

3.5.5 *Training Corpus Dataset Construction*

As shown in Figure 4, the optimized Question Generation Agent θ^* produces new question-context pairs from raw documents. After that, the Answer Generation Agent φ provides answers, resulting in two types of ideal datasets: $\mathcal{I}_{\mathrm{CPT}} = \{(q_j^*, \mathbf{c}_j)\}_{j=1}^{N}$, for continued pre-training (CPT), consisting of question-context pairs. $\mathcal{I}_{\mathrm{SFT}} = \{(q_j^*, \mathbf{c}_j, a_j^*)\}_{j=1}^{N}$, for supervised fine-tuning (SFT), consisting of question-context-answer triples. This collaborative multi-agent pipeline enables scalable, high-quality question answering data construction without reliance on manual annotation.
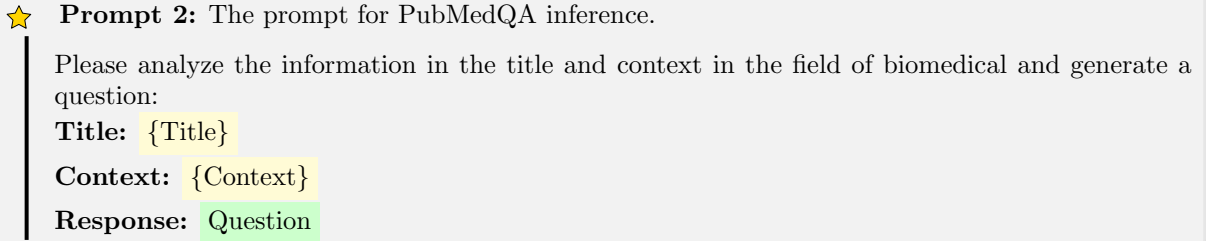
3.6 Training for Downstream Tasks

Finally, the constructed high-quality training corpus is employed to train a target large language model for biomedical question answering in two sequential stages. For instance, we denote the downstream target model as BioMistral [17].

⭐ **Prompt 1**: The prompt for continuous pre-training.

To address the challenges for the biomedical field, I reviewed the paper titled: {Title} , {Context} .

Motivated by this study, I conducted a literature review to gather additional resources and contextualize its findings. During this process, I identified the following key materials: {Retrived Context} .

Reflecting on these insights, I formulated the following research question: {Question} .

Figure 5 The prompt for continuous pre-training.

> ⭐ **Prompt 2:** The prompt for PubMedQA inference.
>
> Please analyze the information in the title and context in the field of biomedical and generate a question:
>
> **Title:** {Title}
>
> **Context:** {Context}
>
> **Response:** Question

Figure 6 The prompt for PubMedQA inference.

3.6.1 *Continuous Pre-training*

The target LLM is first further pre-trained on the heuristic question-context pairs $\mathcal{I}_{\mathrm{CPT}} = \{(q_j, \mathbf{c}_j)\}_{j=1}^N$ generated by the optimized question generator. This stage aims to improve the model's foundational biomedical knowledge and enhance its ability to understand biomedical contexts and question styles. The prompt of this stage is given as Figure 5.

3.6.2 *Supervised Fine-tuning*

Subsequently, the model is fine-tuned on question-context-answer triples $\mathcal{I}_{\mathrm{SFT}} = \{(q_j, \mathbf{c}_j, a_j)\}_{j=1}^N$, as well as optionally on benchmark QA datasets such as PubMedQA PQA-A [43] training set. As shown in Figure 6, this prompt is designed to match the downstream QA format. This stage explicitly optimizes the model for accurate answer generation and robust question understanding.

3.6.3 *Summary of the Training Procedure*

The two-stage training pipeline ensures that the trained LLM can effectively leverage both the large-scale synthetic data from m-KAILIN and high-quality human-annotated benchmarks, substantially boosting downstream biomedical QA performance. In practice, the continuous pre-training step adopts the standard language modeling objective, while the supervised fine-tuning step minimizes the cross-entropy loss between the model's predicted answers and ground-truth answers.

4 Results

4.1 Experimental Setups

4.1.1 *Base Models*

We utilized the llama-series LLM as base models in our primary experiment, while also incorporating BioMistral [17] as the backend of the Question Generation Agent for building the preference dataset and fine-tuning the question generator. We adopted GPT-4o as the backend of the Answer Generation Agent.

4.1.2 *Baselines*

We conducted a comprehensive evaluation of various open-source models, including LLaMA-2 [60], LLaMA-3 [59], Mistral [61], GLM-4 [62], Qwen2.5 [63], and Gemma [64], Deepseek [65], as well as proprietary models like GPT-4 [57] (with MedPrompt [66,67]), and Med-PaLM [68]. In particular, we focused on models specifically trained for the biomedical domain, such as BioMistral [17], PMC-LLaMA [20], HEAL [51], KAILIN [29] (with Qwen2.5, GLM-4, and LLama-3 as backend), and MMedLM [69], to demonstrate the effectiveness.

4.1.3 *Evaluation Benchmarks and Protocols*

We conduct a comprehensive evaluation using a suite of benchmark subsets derived from PubMedQA [43]. We design a multi-dimensional evaluation protocol that probes complementary aspects of biomedical QA

(a) LLMs of fewer than 13B parameters

(b) LLMs of more than 70B parameters
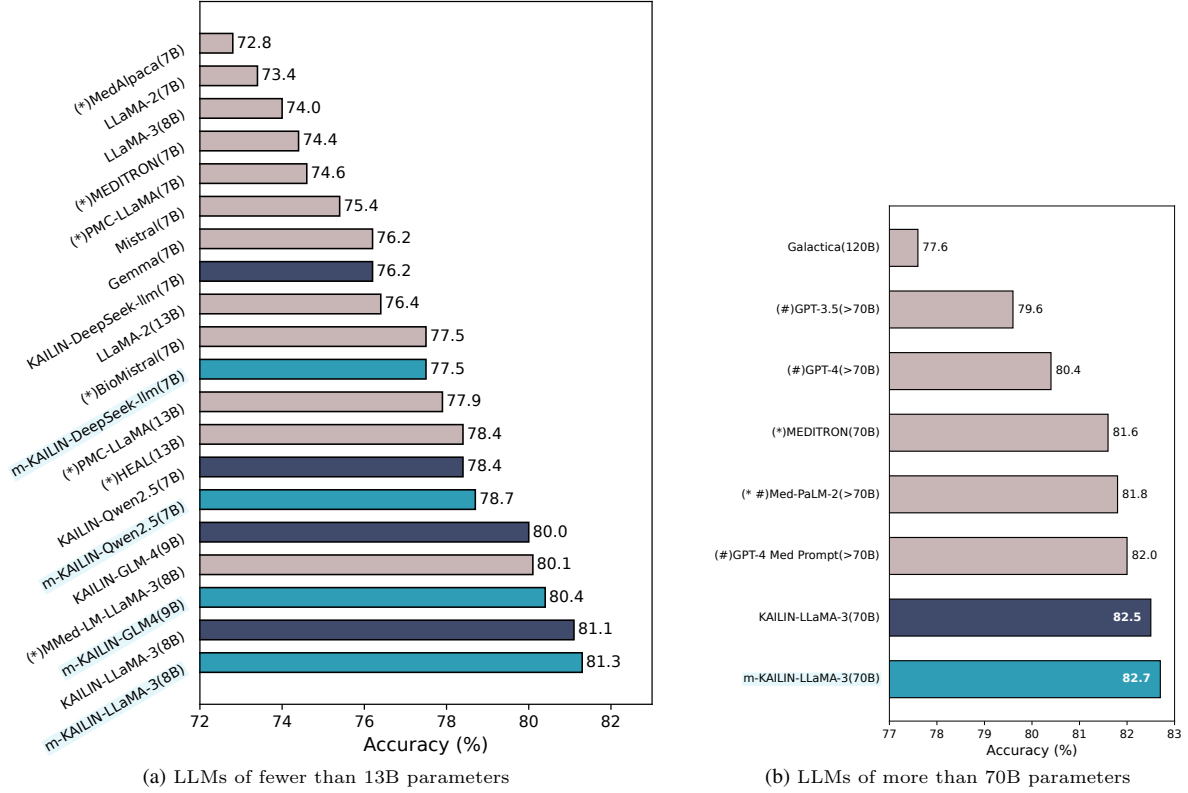
Figure 7 Evaluations (accuracy (%)) for PubMedQA [43] problems on our models compared to other open-source models, closed-source models, and domain-specific models. Models marked with * indicate domain-specific biomedical models. The # symbol denotes closed-source models.

capability, including overall accuracy, temporal generalization, subdisciplinary robustness, reasoning sensitivity, and data scaling behavior. Specifically, we first evaluate all models on the PubMedQA-Full benchmark, which consists of expert-annotated question–answer pairs grounded in biomedical literature and serves as the primary indicator of end-to-end QA performance. To assess robustness under biomedical knowledge evolution, we further construct chronological benchmarks (PubMedQA-Chrono) by partitioning the PQA-L subset into eight time-based slices according to publication periods, enabling evaluation across distinct biomedical eras.

To examine cross-domain generalization, we derive subdisciplinary benchmarks (PubMedQA-MeSH) by grouping samples according to their dominant Medical Subject Headings [55] (MeSH) terms, resulting in six overlapping subsets that reflect major biomedical research areas. In addition, following prior work, we distinguish between reasoning-required and question-only inference settings, forming a reasoning-oriented benchmark that isolates the contribution of contextual understanding and biomedical reasoning.

Together, these benchmark settings provide a controlled yet comprehensive evaluation suite, allowing us to systematically analyze the effectiveness and robustness of the proposed m-KAILIN framework across multiple dimensions of biomedical question answering.

4.1.4 *Platform Information*

Each experiment, either utilizing the API query or performed on an Ubuntu 18.04.6 LTS system featuring an AMD EPYC 7742 processor and eight Cambricon MLU-590B, was carried out within a Python 3.11.0 environment using PyTorch 2.1.1.

4.2 Main Results and Analysis

As illustrated in Figure 7, we benchmarked m-KAILIN and various baseline models across two parameter scales: models with fewer than 13B parameters and those with 70B parameters or greater. A clear positive correlation exists between model size and performance: models with over 70B parameters generally surpass smaller models in accuracy. However, our results reveal that the application of m-KAILIN

narrows this gap, allowing smaller models to achieve competitive performance. Notably, m-KAILIN-LLaMA-3 (8B) nearly matches or outperforms several larger domain-specific models, underscoring the potential of architectural and training innovations to offset limitations in model scale partially. The key driver is the MeSH-based knowledge hierarchy similarity evaluation integrated within Question Evaluation Agent, which substantially enhances a model's ability to comprehend and reason about the complex terminologies and conceptual structures of biomedical texts. This demonstrates the effectiveness of our method in adapting smaller, general-purpose models to specialized domains—a critical advantage given the rapid iteration cycles in large model development. In the large-model group, m-KAILIN significantly boosts the performance of LLaMA-3-70B, enabling it to outperform powerful closed-source systems such as GPT-4 with MedPrompt and Med-PaLM-2, despite their larger parameter counts. This suggests that, while increasing model size typically yields better results, targeted preference alignment using knowledge hierarchy evaluation can enable smaller or open-source models to close the gap or even exceed the performance of much larger models, thus offering substantial efficiency and accessibility benefits. In summary, the m-KAILIN framework enables both small and large general-purpose models to achieve state-of-the-art performance on domain-specific biomedical tasks. These experimental results also validate the feasibility of leveraging the m-KAILIN framework to automatically extract high-quality training data from raw textual sources.

4.3 Ablation Studies of each Technical Component

The main motivation of this ablation study is to systematically dissect the contributions of each key component in our proposed multi-agent framework for biomedical corpus construction. Specifically, we aim to understand how multi-agent diversity (in question generation), domain-specific retrieval, and knowledge-aware evaluation impact the quality of the generated datasets and the downstream performance of LLMs.
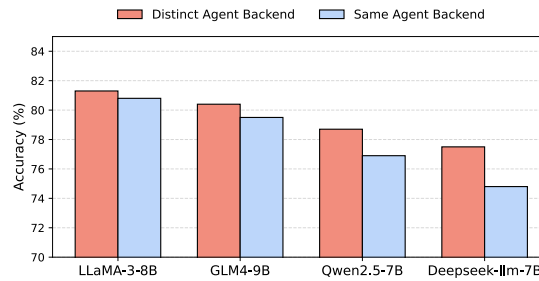
Figure 8 Ablation study on the question generation agent backend. Comparison between using two identical general-purpose agents (Same Agent Backend) and combining a general-purpose agent with a domain-specific agent (Distinct Agent Backend).

4.3.1 *Ablation Study on Question Generation Agent*

We examine whether leveraging heterogeneous agents for question generation—namely, combining one general-purpose and one domain-specific agent (Distinct Agent Backend)—offers advantages over using two identical, general-purpose question generation agents (Same Agent Backend). The rationale is that agent diversity may enhance the richness and coverage of generated questions by introducing complementary knowledge and styles.

As shown in Figure 8, across all evaluated LLMs, the Distinct Agent Backend consistently outperforms the Same Agent Backend. For example, LLaMA-3-8B achieves 81.3% accuracy with the distinct setup, compared to 80.8% with identical agents; similar trends are observed for GLM4-9B, Qwen2.5-7B, and Deepseek-llm-7B. This demonstrates that combining a general and a domain-specific QG agent leads to more diverse and higher-quality question-answer pairs, thereby enhancing downstream QA performance. The results suggest that agent heterogeneity is a key driver for generating a richer and more challenging biomedical corpus.

4.3.2 *Ablation Study on Context Retrieval Agent and Question Evaluation Agent*

To further investigate the roles of domain-specific knowledge and retrieval in our framework, we conduct ablation experiments focusing on the evaluation and retrieval agents. Specifically, we consider two ablation scenarios:
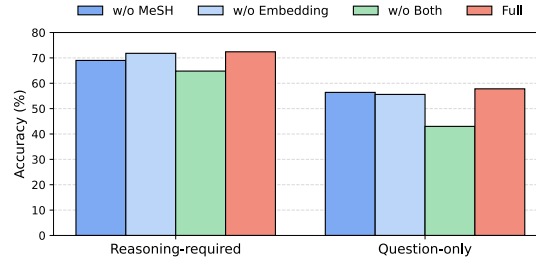
Figure 9 Evaluations of the overall experimental results of the ablation study. The reasoning-required and question-only are both inference settings in PubMedQA study.

• w/o **MeSH**: The evaluation agent does not utilize the MeSH-based knowledge hierarchy; instead, it directly employs a general-purpose LLM for evaluation without hierarchical biomedical guidance.

• w/o **Embedding**: The retrieval agent does not use a domain-adapted embedding model; instead, it relies on a general model such as BERT without further fine-tuning on biomedical data.

Figure 9 shows the performance across different configurations. Removing MeSH (w/o MeSH) leads to a substantial decrease in accuracy, especially in the reasoning-required setting. This suggests that biomedical hierarchical knowledge—provided by MeSH—is crucial for the evaluation agent to accurately judge the relevance and quality of context-question pairs. Without this structured guidance, the general LLM struggles to evaluate biomedical content effectively, resulting in lower overall performance. In the w/o Embedding setting, where the retrieval agent uses general-purpose BERT embeddings instead of a fine-tuned, domain-specific model, performance also drops. This indicates that domain-adapted embeddings significantly enhance the retrieval of relevant biomedical contexts. General embeddings may miss domain-specific nuances, leading to suboptimal context selection and thereby affecting downstream QA accuracy. Notably, when both MeSH and domain-adapted embedding are removed (w/o Both), there is a pronounced performance degradation (64.8% in reasoning-required, 43% in question-only), confirming that both hierarchical knowledge in evaluation and domain adaptation in retrieval play complementary, essential roles. Overall, these results demonstrate that integrating domain-specific components—both in evaluation and retrieval—is vital for constructing a high-quality biomedical corpus and maximizing LLM downstream performance.

4.4 Hyperparameter Experiment

In the m-KAILIN framework, the number of retrieved documents in context, Top-k, is a highly influential parameter. This parameter directly impacts the balance between the comprehensiveness, richness of context information, and the upper limit of contextual understanding of models. We conducted hyperparameter experiments to study the effect of retrieving different numbers of k documents as context on model performance. As shown in Figure 10, the model exhibited the best performance when k was set to 4. This optimal k allows the model to gain a more comprehensive understanding of the overall information relevant to the question in the biomedical domain, improve alignment, and reduce noise. We use this optimal setting for each of the experiments.
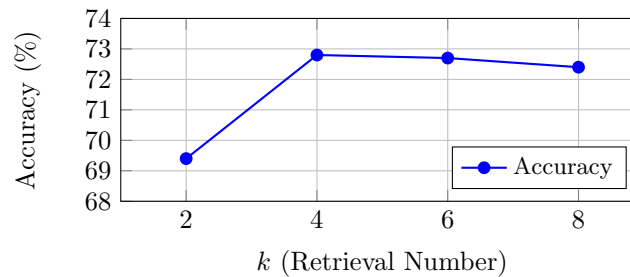
Figure 10 Evaluation(accuracy(%)) on PubMedQA using continuous pre-training datasets of retrieving different Top-k documents.

Figure 11 A case study comparing the generated questions under different experimental settings in the ablation study.

4.5 Case Study of Generated Corpus

From the specific cases illustrated in Figure 11, it is evident that the model can recognize "activity" in relation to papain's function as an enzyme, demonstrating a general understanding facilitated by domain cues. However, without comprehensive and structured biomedical knowledge, the model still exhibits limitations in generating integrative questions that require a deeper or more holistic perspective. Performance differences under the question-only inference setting further highlight these shortcomings. In particular, when the embedding model is not fine-tuned on biomedical data (i.e., using a general-purpose embedding), the model struggles to capture question-specific nuances, resulting in significantly reduced performance when relevant context is unavailable. Our case analysis also reveals the pivotal role of the MeSH-based knowledge hierarchy. When MeSH guidance is present, the model is able to correctly interpret papain as a "proteolytic enzyme" within the broader biomedical context. However, the absence of fine-grained, domain-specific knowledge—especially when leveraging only general LLMs or embeddings—leads to missed opportunities in question generation, such as failing to inquire about detailed aspects like enzyme activity or dimerization. These findings underscore the necessity of integrating both structured biomedical ontologies (e.g., MeSH) and domain-adapted retrieval mechanisms in our multi-agent framework. Only through this combination can the model attain both the breadth and depth required for high-quality, knowledge-rich biomedical question generation and evaluation.

4.6 Scaling Law of Dataset Distillation

We conducted experiments using datasets of varying scales for continuous pre-training, followed by supervised fine-tuning under consistent experimental settings, to evaluate the impact of different data volumes on model performance. In this study, we selected the Llama-3-8B and Llama-3-70B models as base models, to verify the impact of varying dataset scales on models of different sizes.
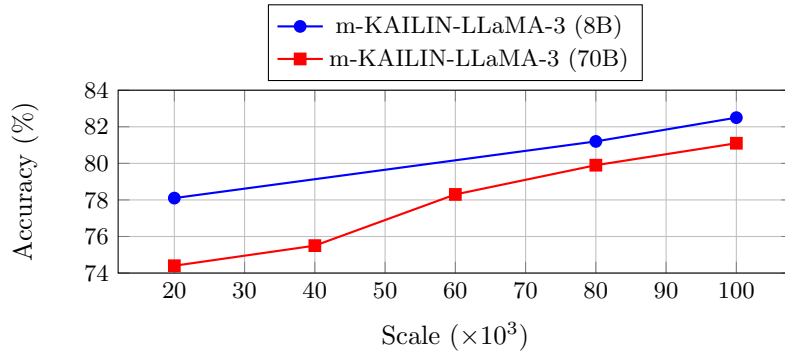

Figure 12 Evaluation (accuracy(%)) of Llama-3-8B and Llama-3-70B on PubMedQA using datasets of varying scales, from 20,000 (20×10^3) to 100,000 (100×10^3).

As shown in Figure 12, we observed that the larger LLaMA-3-70B model demonstrates overall better performance compared to the smaller LLaMA-3-8B model. We attributed this advantage to its larger
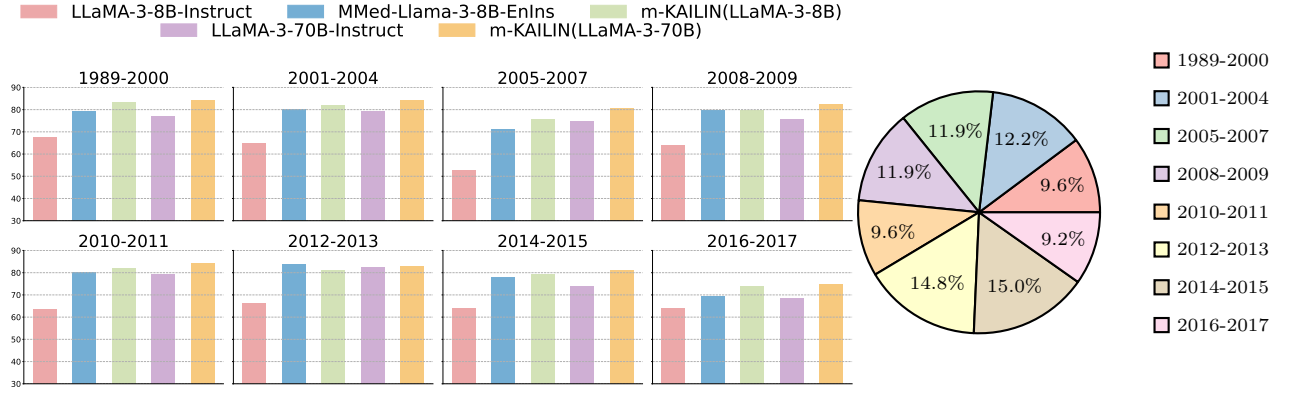
Figure 13 Evaluations (accuracy (%)) for different chronological subsets of problems in PubMedQA PQA-L on our models compared to other representative models.
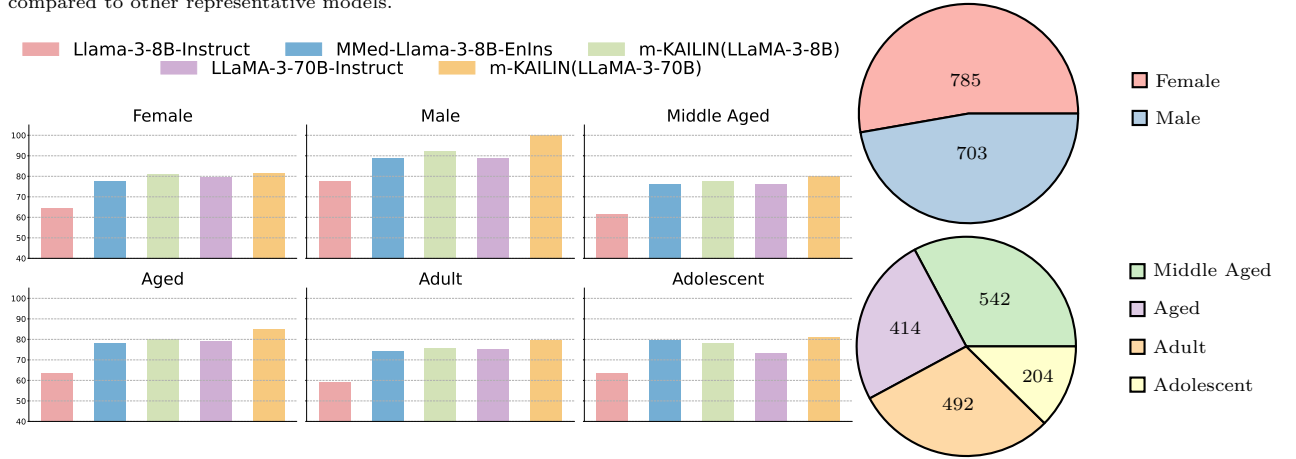
Figure 14 Evaluations (accuracy (%)) for different MeSH term subsets problems in PubMedQA PQA-L on our models compared to other representative models.

parameter scale, which provides superior feature capture and generalization capabilities. We also observed that on datasets of varying sizes, both models showed a trend of performance improvement as the dataset size increased. The underlying driver is that larger datasets distilled by m-KAILIN are better able to represent a broader range of biomedical knowledge due to a greater number of documents being comprehensively analyzed. This phenomenon indicates that the m-KAILIN framework indeed produces higher-quality datasets when distilling larger-scale datasets. Moreover, we found that while smaller models may not outperform larger models, they seem to benefit more from progressively larger distilled datasets. We attribute this phenomenon to the degree of alignment between task complexity and model scale. This phenomenon indicates that larger models, due to potential knowledge redundancy, may gain less from incremental datasets on a specific task.

4.7 Robustness toward Different Question Groups

Our robustness experiments investigate how well the m-KAILIN framework generalizes across different temporal spans and biomedical subdisciplines, reflecting real-world application scenarios that challenge LLMs with both chronological and disciplinary diversity.

4.7.1 *Robustness Under Chronological Question*

As shown in Figure 13, our agentic framework, especially when combined with MeSH-based evaluation, enables models to maintain consistently high accuracy across documents published at different time periods. Compared to strong biomedical baselines such as MMedLM, models distilled with m-KAILIN demonstrate significantly greater robustness to temporal shifts. This improvement stems from two key innovations: (1) the multi-agent setup, which incorporates diverse perspectives during corpus construction, and (2) the MeSH-guided evaluation agent, which provides structured biomedical context, allowing

the model to better understand and adapt to evolving domain knowledge over time. Notably, larger models exhibit even greater temporal robustness, likely due to their increased capacity to absorb and generalize from temporally diverse data curated by our framework.

4.7.2 *Robustness Under Subdisciplinary Question*

We conduct evaluation on subdisciplinary slices based on major MeSH terms (Figure 14). The results show that m-KAILIN-trained models outperform their peers across various biomedical topics and demographics. This robustness stems from our MeSH-based knowledge-driven design, which leverages MeSH hierarchies during both retrieval and evaluation. This design ensures comprehensive coverage and nuanced understanding across biomedical subfields. The evaluation agent's use of MeSH facilitates more relevant and context-aware question-answer generation, systematically exposing the model to the breadth of the biomedical literature's structure. Finally, the overall trend affirms the effectiveness of our approach in producing models well-adapted to the complex, hierarchical landscape of biomedical knowledge.

5 Conclusion

In this work, we present a knowledge-driven, multi-agent framework for scientific corpus distillation tailored to biomedical large language model training. By leveraging a collaborative architecture—where specialized agents guided by biomedical ontologies autonomously generate, evaluate, and refine question-answer pairs—our approach addresses the limitations of existing open-source scientific corpora in both scale and quality. Through extensive experiments, we demonstrate that language models trained on our multi-agent distilled datasets achieve substantial improvements in biomedical question-answering tasks, outperforming both strong open-source and proprietary baselines. Our ablation studies further validate the effectiveness and synergy of each agent within the framework. This study highlights the potential of agentic, knowledge-guided corpus construction for advancing biomedical AI, and provides scalable tools and datasets to the community for future research.

References

1   Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

2   Yi Luo, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu, and Chen Lin. From intention to implementation: automating biomedical research via llms. *Science China Information Sciences*, 68(7):1–18, 2025.

3   Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263, 2023.

4   Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

5   Wei Lan, Zhentao Tang, Mingyang Liu, Qingfeng Chen, Wei Peng, Yiping Phoebe Chen, and Yi Pan. The large language models on biomedical data analysis: A survey. *IEEE Journal of Biomedical and Health Informatics*, pages 1–13, 2025.

6   Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

7   Chengrui Wang, Qingqing Long, Meng Xiao, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. Biorag: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107*, 2024.

8   Zhiwen Xie, Zhao Zeng, Guangyou Zhou, and Weijun Wang. Topic enhanced deep structured semantic models for knowledge base question answering. *Science China Information Sciences*, 60(11):110103, 2017.

9   Vedansh Thakkar, Greg M Silverman, Abhinab Kc, Nicholas E Ingraham, Emma Jones, Samantha King, and Christopher J Tignanelli. Comparison of large language models versus traditional information extraction methods for real world evidence of patient symptomatology in acute and post-acute sequelae of sars-cov-2. 2024.

10   Yiming Lin, Madelon Hulsebos, Ruiying Ma, Shreya Shankar, Sepanta Zeighami, Aditya G Parameswaran, and Eugene Wu. Querying templatized document collections with large language models. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 2422–2435. IEEE Computer Society, 2025.

11   Chuan Qin, Xin Chen, Chengrui Wang, Pengmin Wu, Xi Chen, Yihang Cheng, Jingyi Zhao, Meng Xiao, Xiangchao Dong, Qingqing Long, et al. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. *arXiv preprint arXiv:2503.13503*, 2025.

12   Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI, 2024.

13   Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.

14   Yihao Ai, Zhiyuan Ning, Weiwei Dai, Pengfei Wang, Yi Du, Wenjuan Cui, Kunpeng Liu, and Yuanchun Zhou. Distilling closed-source llm's knowledge for locally stable and economic biomedical entity linking. 2025.

15   Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, et al. A survey for large language models in biomedicine. *arXiv preprint arXiv:2409.00133*, 2024.

16   Sami Barrit, Salim El Hadwe, Romain Carron, and Joseph R Madsen. Rise of large language models in neurosurgery. *Journal of Neurosurgery*, 1(aop):1–2, 2024.

17 Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

18 Philipp Angerer, Lukas Simon, Sophie Tritschler, F Alexander Wolf, David Fischer, and Fabian J Theis. Single cells make big data: new challenges and opportunities in transcriptomics. *Current opinion in systems biology*, 4:85–91, 2017.

19 Ling Yang, Jun-Hai Yong, Hongzhi Yin, Jiawei Jiang, Meng Xiao, Wentao Zhang, Bin Wang, et al. Distribution-aware data expansion with diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

20 Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

21 Jianbo Yuan, Zhiwei Jin, Han Guo, Hongxia Jin, Xianchao Zhang, Tristram Smith, and Jiebo Luo. Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowledge and Information Systems*, 62:317–336, 2020.

22 Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2024.

23 Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

24 Xunxin Cai, Meng Xiao, Zhiyuan Ning, and Yuanchun Zhou. Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 956–961. IEEE, 2023.

25 Tianqi He, Xiaohan Huang, Yi Du, Qingqing Long, Ziyue Qiao, Min Wu, Yanjie Fu, Yuanchun Zhou, and Meng Xiao. Fastft: Accelerating reinforced feature transformation via advanced exploration strategies. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 4120–4133. IEEE Computer Society, 2025.

26 Meng Xiao, Min Wu, Ziyue Qiao, Yanjie Fu, Zhiyuan Ning, Yi Du, and Yuanchun Zhou. Interdisciplinary fairness in imbalanced research proposal topic inference: A hierarchical transformer-based method with selective interpolation. *ACM Transactions on Knowledge Discovery from Data*, 19(2):1–21, 2025.

27 Meng Xiao, Ziyue Qiao, Yanjie Fu, Hao Dong, Yi Du, Pengyang Wang, Hui Xiong, and Yuanchun Zhou. Hierarchical interdisciplinary topic detection model for research proposal classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9685–9699, 2023.

28 Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*, 2023.

29 Xunxin Cai, Chengrui Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. Knowledge hierarchy guided biological-medical dataset distillation for domain llm training. *arXiv preprint arXiv:2501.15108*, 2025.

30 Meng Xiao, Ziyue Qiao, Yanjie Fu, Yi Du, Pengyang Wang, and Yuanchun Zhou. Expert knowledge-guided length-variant hierarchical label generation for proposal classification. In *2021 ieee international conference on data mining (icdm)*, pages 757–766. IEEE, 2021.

31 Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

32 Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

33 Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *arXiv preprint arXiv:2504.14772*, 2025.

34 Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

35 Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022.

36 Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. *arXiv preprint arXiv:2011.00050*, 2020.

37 Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, 34:5186–5198, 2021.

38 Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

39 Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, pages 12352–12364. PMLR, 2022.

40 Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z Pan. Delving into effective gradient matching for dataset condensation. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE, 2023.

41 George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022.

42 Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023.

43 Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

44 Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. Improving small language models on pubmedqa via generative data augmentation. *arXiv preprint arXiv:2305.07804*, 2023.

45 Inthrani Raja Indran, Priya Paramanathan, Neelima Gupta, and Nurulhuda Mustafa. Twelve tips to leverage ai for efficient and effective medical question generation: A guide for educators using chat gpt. *Medical Teacher*, pages 1–6, 2023.

46 Yuqi Wu, Kaining Mao, Yanbo Zhang, and Jie Chen. Callm: Enhancing clinical interview analysis through data augmentation with large language models. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7531–7542, 2024.

47 Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2024.

48 Yichen Gao, Licheng Zong, and Yu Li. Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation. *CLEF Working Notes*, 2024.

49 Jan Heinrich Merker, Alexander Bondarenko, Matthias Hagen, and Adrian Viehweger. Mibi at bioasq 2024: retrieval-augmented generation for answering biomedical questions. 2024.

50 Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, page 104662, 2024.

51 Dong Yuan, Eti Rastogi, Gautam Naik, Jai Chintagunta, Sree Prasanna Rajagopal, Fen Zhao, Sagar Goyal, and Jeff Ward.

A continued pretrained llm approach for automatic medical note generation. *arXiv preprint arXiv:2403.09057*, 2024.

52 Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.

53 Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

54 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

55 Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

56 Dekang Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304, 1998.

57 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

58 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

59 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

60 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

61 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

62 Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

63 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

64 Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

65 Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

66 Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

67 Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

68 Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

69 Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*, 2024.