Knowledge Hierarchy Guided Biological-Medical Dataset Distillation for Domain LLM Training

Xunxin Cai[1,2†], Chengrui Wang[1†], Qingqing Long[1], Yuanchun Zhou[1,2,3], and Meng Xiao[1(✉)]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
{xxcai,crwang,qqlong,zyc,shaow}@cnic.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences

Abstract. The rapid advancement of large language models (LLMs) in biological-medical applications has highlighted a gap between their potential and the limited scale and often low quality of available open-source annotated textual datasets. In addition, the inherent complexity of the biomedical knowledge hierarchy significantly hampers efforts to bridge this gap. Can LLMs themselves play a pivotal role in overcoming this limitation? Motivated by this question, we investigate this challenge in the present study. We propose a framework that automates the distillation of high-quality textual training data from the extensive scientific literature. Our approach self-evaluates and generates questions that are more closely aligned with the biomedical domain, guided by the biomedical knowledge hierarchy through medical subject headings (MeSH). This comprehensive framework establishes an automated workflow, thereby eliminating the need for manual intervention. Furthermore, we conducted comprehensive experiments to evaluate the impact of our framework-generated data on downstream language models of varying sizes. Our approach substantially improves question-answering tasks compared to pre-trained models from the life sciences domain and powerful close-source models represented by GPT-4. Notably, the generated AI-Ready dataset enabled the Llama3-70B base model to outperform GPT-4 using MedPrompt with multiple times the number of parameters. Detailed case studies and ablation experiments underscore the significance of each component within our framework[4].

1 Introduction

The rise of LLMs has revolutionized bioinformatics, driving the adoption of automated applications across areas [22], with their effectiveness increasingly validated in real-world Question-Answer (QA) [11]. Nevertheless, the inherent complexity of biomedical tasks means that general LLMs often fail to give correct answer unless they are carefully adapted and fine-tuned [21]. Additionally, the limited availability of substantial biomedical text data impedes the fine-tuning

[†] These authors contributed equally to this work.
[4] Our code is shared on Github: link.

of domain-specific LLMs. While biomedical research papers indeed serve as rich sources of quality and dependable corpora, they are characterized by complex terminologies and detailed conceptual frameworks that demand considerable human effort for understanding and processing. [25,3]. Those observations lead to an essential question: **How to automatically distill high-quality, large-scale datasets from extensive research papers, thus support LLM training?**

To address the automated corpora distillation challenge, existing frameworks can be categorized primarily into three approaches: (1) *Predefined rules-based approaches* [9,2] undertake extensive data cleaning by filtering and standardizing large-scale bioinformatics datasets. While those approaches reduce noise and improve data quality, they incur significant operational costs and limit scalability due to the human labor. (2) *Knowledge graph-based approaches* [26,31] leverage biomedical text data to create comprehensive knowledge structures, but the reliance on curated databases results in inefficiencies and scalability challenges. (3) *Synthesis approaches* [8,19,4] present a promising automated solution to generate question-answer pairs and process large volumes of documents by using LLMs. Nonetheless, these studies neglect the integration of cross-disciplinary collaboration [29,10], resulting in a lack of diversity and reliability.

Motivated by those limitations, we propose **K**nowledge hier**A**rchy gu**I**ded biologica**L**-med**I**cal dataset distillatio**N** (**KAILIN**), an automated framework that integrates knowledge hierarchy and utilizes multiple LLMs as experts for domain QA training corpora extraction. The core idea of KAILIN is to introduce a well-established knowledge hierarchy (i.e., Medical Subject Headings (MeSH) [13]) to assess the alignment of the generated 'Question-Answer-Context' pair to the domain understanding. This framework begins with fine-tuning two LLMs to generate questions from annotated yet scarce open-source datasets. After that, the framework retrieves the context from 23 million collected research articles that are most related to the generated questions. The better question is determined by evaluating the retrieved contexts and selecting the one with the superior alignment score to the knowledge hierarchy. By that, the automated pipeline of generating preference data is present. This dataset is designed to train a language model to craft improved questions from unannotated research articles that align more effectively with the existing knowledge structure. Using the improved question, we can retrieve the related context, generate answers, and finally form the AI-Ready dataset.

In summary, the key contributions of this work can be summarized as:

- **Biomedical Dataset Distillation Workflow:** We present a comprehensive, highly automated workflow for distilling biomedical corpora from large-scale research articles. This framework enables the creation of expansive, domain-specific training datasets without the need for manual annotation, significantly reducing the cost and time involved in dataset preparation.
- **Framework and Methodology:** We proposed the **KAILIN** framework which incorporates a MeSH-based knowledge hierarchy similarity evaluation method to integrate and evaluate the quality of the distilled biomedical corpora. KAILIN efficiently constructs high-quality datasets by combining

knowledge-based evaluation with context-aware selection, obviating the need for human intervention in dataset curation.

- **Empirical Validation and Insights:** We conducted extensive experiments to validate the effectiveness of our framework and the resulting datasets. Through ablation studies and case analyses, we explored the impact of each technical component. Additionally, we investigated the scaling law of dataset distillation across various settings and model hyperparameter selections.

2 Methodology

In this section, we introduce the KAILIN framework, which aims to enhance the open-source dataset through the dataset distillation process, and validate the effectiveness on downstream tasks.

Fine-Tuning Question Generator: To enhance the performance of general-purpose base models in biomedical question generation, we employed biomedical open-source BioASQ dataset[24] as the training set $\mathcal{T}$. Using this training set $\mathcal{T}$, we trained two distinct question generators θ^1 and θ^2, with LLaMA-2-7B and BioMistral as base models θ respectively.

Retrieval Process: We collected 23 million abstracts from PubMed[5], which serves as the raw dataset $\mathcal{R}$ for dataset distillation and is also constructed as a vector database $\mathcal{V}$ for retrieval purposes. Given a document $d_i \in \mathcal{R}$, and two question generation models θ^1 and θ^2 that use d_i as input to infer questions q_i^1 and q_i^2. We then used these questions q_i^1 and q_i^2 respectively as input to retrieve the top-k most similar documents from the vector database $\mathcal{V}$. These retrieved raw documents, serving as the contexts c_i^1 and c_i^2 associated with q_i^1 and q_i^2 respectively, will be used for the subsequent knowledge hierarchy similarity evaluation.

Knowledge Hierarchy Similarity Evaluation: Using the retrieved contexts c_i^1 and c_i^2 associated with the generated questions q_i^1 and q_i^2, we conducted a knowledge hierarchy similarity evaluation between each context and the original document d_i that used for generating the questions. This allows us to select the question that better aligns with the knowledge hierarchy of the biomedical field. For similarity evaluation, we introduced Medical Subject Headings (MeSH), denoted as $\mathcal{M}$. MeSH is a hierarchical classification system centered on an well-organized vocabulary that systematically classifies and organizes biomedical knowledge through structured subject terms. We present $\prec$, a partial order illustrating the *Belong-to* relationship, to explain how various terms are interrelated. The characteristics of $\prec$ include being asymmetric, anti-reflexive, and transitive[27,28]:

[5] PubMed:`https://pubmed.ncbi.nlm.nih.gov/`

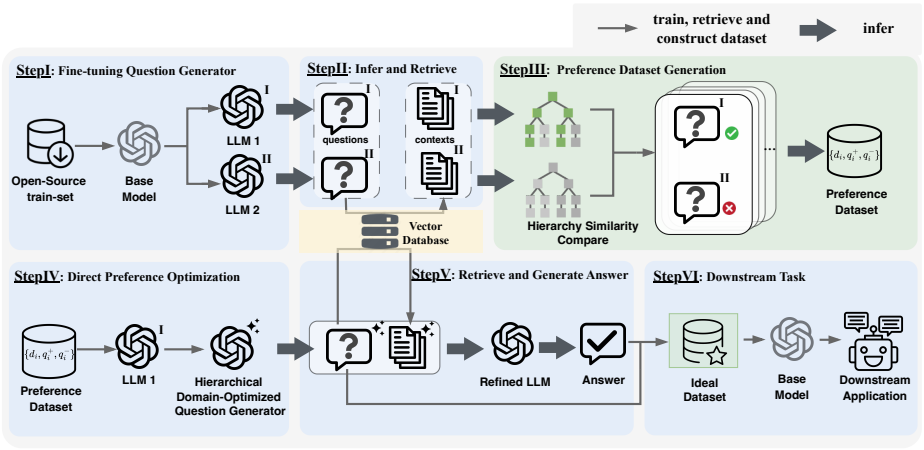


Fig. 1: The overview of KAILIN framework.

- The only one greatest category *root* is the root of the $\mathcal{M}$,
- $\forall m_i^x \in M_i, m_j^y \in M_j, m_i^x \prec m_j^y \rightarrow m_j^y \not\prec m_i^x$,
- $\forall m_i^x \in M_i, m_i^x \not\prec m_i^x$,
- $\forall m_i^x \in M_i, m_j^y \in M_j, m_k^z \in M_k, m_i^x \prec m_j^y \wedge m_j^y \prec m_k^z \rightarrow m_i^x \prec m_k^z$.

Finally, we define the Hierarchical MeSH Structure $\mathcal{M}$ as a partial order set $\mathcal{M} = (\mathbf{M}, \prec)$, where $\mathbf{M} = \{M_i\}_{i=1}^n$ is a level-organized term set and n denote the total depth. We then analyzed the structured subject terms in the contexts c_i^1, c_i^2 and the original document d_i, incorporating the information content of their hierarchical positions. For any structured subject term m, we first calculate its information content as:

$$IC(m) = -\log(\frac{freq(\mathcal{M}(m))}{n_{terms}}), \tag{1}$$

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. The IC reflects the specificity of a term; rarer terms have higher IC values. For any two MeSH terms m^x and m^y, we identify their Lowest Common Ancestor (LCA) in the MeSH hierarchy as $\Lambda(m^x, m^y)$. Referring to Lin's approach[12], we calculate the semantic similarity in a taxonomy based on information content as:

$$S_{x,y} = \frac{2 \times IC(\Lambda(m^x, m^y))}{IC(m^x) + IC(m^y)}, \tag{2}$$

where the $S_{x,y}$ represent similarity between m^x and m^y. We then calculate the final similarity of the knowledge hierarchy between context c_i^1 and original document d_i by averaging over all pairwise comparisons between terms as:

$$\bar{S}_i^1 = \frac{1}{|d_i||c_i^1|} \sum_{m^x \in d_i} \sum_{m^y \in c_i^1} S_{x,y}, \tag{3}$$

where $\bar{S}_i^1$ denotes the knowledge hierarchy similarity between original document d_i and associated context c_i^1 for question q_i^1. Similarly, we calculate the knowledge hierarchy similarity $\bar{S}_i^2$ between the original document d_i and c_i^2.

Preference Dataset Construction. Using the knowledge hierarchy similarity evaluation metrics, we conducted a similarity comparison of the generated questions q_i^1 and q_i^2 on a large number of original documents $d_i \in \mathcal{R}$. By comparing the corresponding similarity $\bar{S}_i^1$ and $\bar{S}_i^2$, we assessed the alignment of q_i^1 and q_i^2 with the biomedical knowledge hierarchy. Based on such comparison, we consider q_i^1 or q_i^2 with the higher similarity score to be better aligned and designate it as q_i^+, while the other is designated as q_i^- and constructed a preference dataset $\mathcal{P} = \{d_i, q_i^+, q_i^-\}_{i=1}^N$.

Direct Preference Optimization. With the preference data pairs prepared offline as $\mathcal{P}$, we employed direct preference optimization (DPO)[17] for model alignment. We performed DPO to get optimized question generator θ_3, $\theta_1 \xrightarrow{\mathcal{P}} \theta^3$.

Ideal Dataset Construction We employed the optimized question generation model θ_3 to utilize a original document $d_j \in \mathcal{R}$ as input and generate optimized question q_j. Documents related to q_j were then retrieved as context c_j, and we further employed the LLaMA-3-70B and GPT-4o to generate ideal answers a_j. We constructed two different ideal datasets for distinct purposes. To enhance the model's foundational understanding in the biomedical field, we combined questions q_j and relevant contexts c_j to form an ideal dataset $\mathcal{I}_1 = \{q_j, c_j\}_{j=1}^N$ for continued pre-training. For improving question-answering performance, we combined questions q_j, relevant contexts c_j, and corresponding answers a_j to form an ideal supervised fine-tuning dataset $\mathcal{I}_2 = \{q_j, c_j, a_j\}_{j=1}^N$.

Training for Downstream Task When we proceeded with further training for biomedical question-answering applications, we utilized the $\mathcal{I}_1$ dataset for continued pretraining to achieve better performance improvements along with the subsequent supervised fine-tuning. We describe the two training stages in detail as:

- **Continuous Pre-training.** We utilize heuristic questions q_j generated by the KAILIN framework, along with the retrieved Top-k documents c_j associated with them, as the corpus for continuous pre-training, using the prompt. With the pre-training corpus constructed in this manner, we continued to pre-train our base models with varying parameter sizes.

- **Supervised Fine-tuning.** To optimize the model's question-answering performance following continuous pre-training, we further performed full-parameter fine-tuning on the models with the PQA-A training set from PubMedQA, with the prompt.

3 Experiment

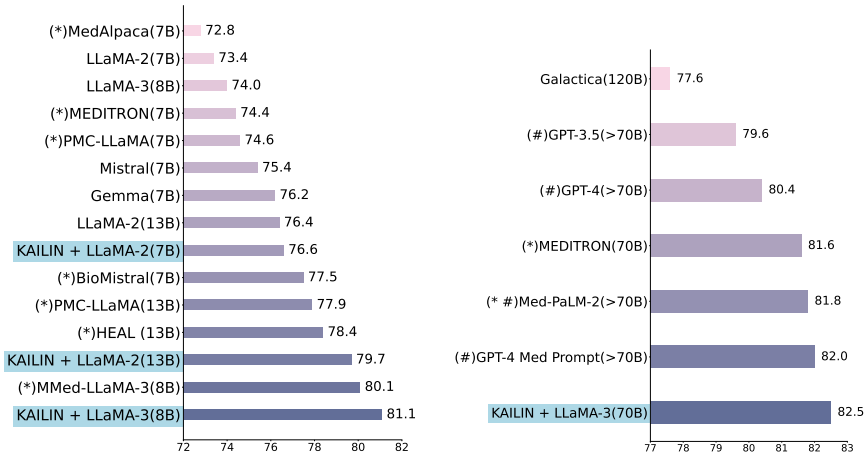
3.1 Experimental Setups

Base Models and Baselines. We utilized Llama-2-7B, Llama-2-13B [23], Llama-3-8B, and Llama-3-70B [5] as the base models in our primary experiment, while also incorporating BioMistral [9] for building the preference dataset in training the question generator. We conducted a comprehensive evaluation of various open-source models, including LLaMA-2 [23], LLaMA-3 [5], Mistral [6], and Gemma[20], as well as proprietary models like GPT-4[1], and PaLM [18]. In particular, we focused on models specifically trained for the biomedical domain, such as BioMistral [9], PMC-LLaMA [26], HEAL [30], and MMedLM [16], to demonstrate the effectiveness of our approach.

Evaluation Datasets. We validated the results of our main experiment using the PubMedQA benchmark [7], a dataset specifically designed to assess the performance of question-answering systems in the biomedical domain. PubMedQA is tailored to address questions relevant to biomedical literature, making it highly suitable for assessing our framework's adaption in this field. Additionally, we categorized the benchmark based on Medical Subject Headings (MeSH) [13] and publication dates, enabling us to evaluate our system's improved understanding and robustness across diverse biomedical topics and varying time spans.

3.2 Main Results and Analysis

We compared KAILIN against various models divided into two groups based on parameter size: those with fewer than 13B parameters and those with 70B parameters or more. When comparing domain-specific models with general-purpose models within LLMs with fewer than 13B parameters, as shown in Figure 2, we observed that general-purpose models tend to struggle more to excel in domain-specific tasks. However, the KAILIN framework enables general-purpose models to outperform domain-specific models with higher training costs, even with minimal additional training. The underlying driver stems from the KAILIN framework's use of MeSH-based knowledge hierarchy similarity evaluation, which effectively addresses the comprehension challenges posed by the rich terminologies and complex conceptual structures inherent in biomedical texts. This phenomenon indicates that the KAILIN framework excels in training small general-purpose models to adapt more effectively to specific domains, which is particularly advantageous in the fast-paced evolution of large model iterations.



(a) LLMs of fewer than 13B parameters (b) LLMs of more than 70B parameters

Fig. 2: Evaluations (accuracy (%)) for PubMedQA [7] problems on our models compared to other open-source models, closed-source models, and domain-specific models. Models marked with * indicate that they are domain-specific large models focused on the biomedical field rather than general-purpose models. The # symbol denotes closed-source models as opposed to open-source ones.

When comparing LLMs with more than 70B parameters, we observed that the KAILIN framework enables LLaMA-3-70B to outperform closed-source models with significantly larger parameter counts, such as GPT-4 with MedPrompt[15,14] and Med-PaLM-2[18]. While larger models typically demonstrate superior performance, the KAILIN framework leverages MeSH-based knowledge hierarchy similarity evaluation for preference alignment. This approach acts as a pivotal underlying driver, enabling the model to excel in specific tasks and surpass significantly larger counterparts. This phenomenon also highlights a potential future direction for training domain-specific models: KAILIN demonstrates the use of smaller datasets that retain a comprehensive understanding of the knowledge hierarchy to optimize performance on domain-specific tasks.

3.3 Ablation Study

We investigated the impact of our MeSH-based knowledge hierarchy similarity evaluation by instead utilizing a Term Frequency-Inverse Document Frequency (TF-IDF) approach for evaluating the document collection. We also investigated the influence of the embedding model during retrieval process by randomly sampling top-k documents and analyzing their replacement in the MeSH-based preference selection process.

As shown in Table 1, we found that the overall performance was worse under the ablation setting without MeSH than without Embedding model, highlighting

Table 1: Evaluations (accuracy (%)) of the overall experimental results of the ablation study. The reasoning-required and question-only are both inference settings in PubMedQA study.

	Reasoning-required	**Question-only**
w/o **MeSH**	69	56.4
w/o **Embedding**	71.8	55.6
w/o **Both**	64.8	43
Full	**72.4**	**57.8**

the critical role of MeSH as part of the framework. Without MeSH serving as a structured marker for each document within the knowledge hierarchy, the model struggled to align with the overall knowledge hierarchy of the biomedical field through isolated documents.

4 Conclusion

In this paper, we proposed a novel automated dataset distillation framework, namely KAILIN, which integrates domain-specific knowledge hierarchy. We aim for KAILIN to serve as an automated approach that distlls datasets of high-quality from existing unlabeled datasets at lower costs while preserving the integrity of domain knowledge hierarchy. Our method was evaluated on the PubMedQA leaderboard, and we further assessed performance robustness across subsets divided by time span and disciplines. These evaluations revealed how LLMs perform differently across various time periods and subfields, demonstrating the superiority of our approach.

5 Ackownledgement

This work is partially supported by the Beijing Natural Science Foundation (No.4254089), the Postdoctoral Fellowship Program of CPSF (No.GZC20232736), the China Postdoctoral Science Foundation Funded Project (No.2023M743565), and National Natural Science Foundation of China (No.92470204).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Angerer, P., Simon, L., Tritschler, S., Wolf, F.A., Fischer, D., Theis, F.J.: Single cells make big data: new challenges and opportunities in transcriptomics. Current opinion in systems biology **4**, 85–91 (2017)
3. Barrit, S., El Hadwe, S., Carron, R., Madsen, J.R.: Rise of large language models in neurosurgery. Journal of Neurosurgery **1**(aop), 1–2 (2024)

4. Cai, X., Xiao, M., Ning, Z., Zhou, Y.: Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In: 2023 IEEE International Conference on Data Mining (ICDM). pp. 956–961. IEEE (2023)
5. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
6. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
7. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146 (2019)
8. Kumichev, G., Blinov, P., Kuzkina, Y., Goncharov, V., Zubkova, G., Zenovkin, N., Goncharov, A., Savchenko, A.: Medsyn: Llm-based synthetic medical text generation framework. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 215–230. Springer (2024)
9. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)
10. Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations. arXiv preprint arXiv:2310.07849 (2023)
11. Liévin, V., Hother, C.E., Motzfeldt, A.G., Winther, O.: Can large language models reason about medical questions? Patterns **5**(3) (2024)
12. Lin, D., et al.: An information-theoretic definition of similarity. In: Icml. vol. 98, pp. 296–304 (1998)
13. Lipscomb, C.E.: Medical subject headings (mesh). Bulletin of the Medical Library Association **88**(3), 265 (2000)
14. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023)
15. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al.: Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv preprint arXiv:2311.16452 (2023)
16. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., Xie, W.: Towards building multilingual language model for medicine. arXiv preprint arXiv:2402.13963 (2024)
17. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36** (2024)
18. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
19. Tang, R., Han, X., Jiang, X., Hu, X.: Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360 (2023)
20. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
21. Thakkar, V., Silverman, G.M., Kc, A., Ingraham, N.E., Jones, E., King, S., Tignanelli, C.J.: 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)

22. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)
23. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
24. Tsatsaronis, G., Schroeder, M., Paliouras, G., Almirantis, Y., Androutsopoulos, I., Gaussier, E., Gallinari, P., Artieres, T., Alvers, M.R., Zschunke, M., et al.: Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In: 2012 AAAI Fall Symposium Series (2012)
25. Wang, C., Li, M., He, J., Wang, Z., Darzi, E., Chen, Z., Ye, J., Li, T., Su, Y., Ke, J., et al.: A survey for large language models in biomedicine. arXiv preprint arXiv:2409.00133 (2024)
26. Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., Wang, Y.: Pmc-llama: toward building open-source language models for medicine. Journal of the American Medical Informatics Association p. ocae045 (2024)
27. Wu, F., Zhang, J., Honavar, V.: Learning classifiers using hierarchically structured class taxonomies. In: International symposium on abstraction, reformulation, and approximation. pp. 313–320. Springer (2005)
28. Xiao, M., Qiao, Z., Fu, Y., Dong, H., Du, Y., Wang, P., Xiong, H., Zhou, Y.: Hierarchical interdisciplinary topic detection model for research proposal classification. IEEE Transactions on Knowledge and Data Engineering **35**(9), 9685–9699 (2023). https://doi.org/10.1109/TKDE.2023.3248608
29. Xiao, M., Wu, M., Qiao, Z., Fu, Y., Ning, Z., Du, Y., Zhou, Y.: Interdisciplinary fairness in imbalanced research proposal topic inference: A hierarchical transformer-based method with selective interpolation. ACM Transactions on Knowledge Discovery from Data
30. Yuan, D., Rastogi, E., Naik, G., Chintagunta, J., Rajagopal, S.P., Zhao, F., Goyal, S., Ward, J.: A continued pretrained llm approach for automatic medical note generation. arXiv preprint arXiv:2403.09057 (2024)
31. Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., Luo, J.: Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowledge and Information Systems **62**, 317–336 (2020)