Applied Soft Computing Journal 175 (2025) 113084

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Efficient fine-tuning of small-parameter large language models for biomedical bilingual multi-task applications

Yinghong Li [a,*,1], Yudong Yan [a,1], Zhuohao Tong [a], Yu Wang [a], Yinqi Yang [a], Mingze Bai [a], Dan Pu [a], Jiazheng Xie [a], Chuan Liu [a], Bo Li [b,*], Mingwei Liu [c], Kunxian Shu [a,*]

[a] Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, PR China
[b] College of Life Sciences, Chongqing Normal University, Chongqing 401331, PR China
[c] Key Laboratory of Clinical Laboratory Diagnostics, College of Laboratory Medicine, Chongqing Medical University, Chongqing 400016, PR China

HIGHLIGHTS

- BioQwen: a series of small-parameter, high-performance models for biomedical bilingual multi-task applications.
- Innovative training methods: a two-stage fine-tuning strategy enhanced by QLoRA and FlashAttention-2.
- Developed a novel data preprocessing pipeline, which was confirmed to be effective by evaluation.

ARTICLE INFO

Keywords:
Biomedical NLP
Chinese-English bilingual text mining
Fine-tuning
Multi-task models
Small-parameter LLMs

ABSTRACT

The escalating computational costs of large language models (LLMs) have catalyzed the pursuit of more efficient alternatives, particularly in specialized domains like biomedicine. In this study, we propose BioQwen, a series of small-parameter biomedical bilingual (Chinese–English) multi-task models designed to mitigate the resource demands of LLMs while achieving high performance.

BioQwen is trained on carefully curated open-source biomedical datasets, employing a stringent preprocessing pipeline with thorough quality filtering and standardized formatting. Through an efficient two-stage fine-tuning strategy, BioQwen models with 0.5B, 1.5B, and 1.8B parameters attain competitive performance across a variety of comprehension and generative tasks. For comprehension tasks, BioQwen-1.8B achieves a Macro F1 score of 0.730 and a balanced accuracy of 0.802 on the BC5CDR dataset, surpassing the 7B-parameter Taiyi model's scores of 0.685 and 0.757. In generative tasks, BioQwen delivers superior zero-shot results on the iCliniq dataset, outperforming all baselines across multiple metrics. Comparisons with established small-parameter LLMs (e.g., Llama3.2 1B) further substantiate the effectiveness of domain-specific fine-tuning.

Significantly, BioQwen's reduced iteration time highlights its computational efficiency, and its quantized version demonstrates successful deployment on mobile devices, confirming its viability in resource-constrained settings. This study demonstrates the potential of strategically fine-tuned small-parameter LLMs to deliver resource-efficient, high-performing solutions for biomedical bilingual applications, expanding accessibility and usability in the field.

1. Introduction

Large language models (LLMs) have demonstrated exceptional performance across diverse natural language processing (NLP) tasks, revolutionizing the field with their ability to generalize and tackle complex challenges.In biomedical fields such as clinical research [1] and medical education [2], these models have proven their potential application value. Notably, OpenAI's ChatGPT has sparked significant influence in both academia and industry, greatly advancing NLP technology [3]. According to research by Jared Kaplan et al. [4], model performance appears to be proportional to their number of parameters, as evidenced by the release of larger and more powerful models like GPT-4.

* Corresponding authors.
E-mail addresses: liyinghong@cqupt.edu.cn (Y. Li), libcell@cqnu.edu.cn (B. Li), shukx@cqupt.edu.cn (K. Shu).
[1] These authors contributed equally: Yinghong Li and Yudong Yan

https://doi.org/10.1016/j.asoc.2025.113084
Received 18 July 2024; Received in revised form 9 January 2025; Accepted 25 March 2025
Available online 28 March 2025

Over the past few years, researchers worldwide have made tremendous efforts to scale up models and datasets, significantly driving the development of artificial intelligence technology [5]. The number of model parameters has increased from 1.5 billion in GPT-2 [6] five years ago to 175 billion in GPT-3.5-turbo [7], and further expanded to 314 billion in the largest current open-source model, Grok-1 [8], showing a continuous growth trend. The rapid increase in model parameters, along with complex training processes such as continual pre-training and reinforcement learning from human feedback (RLHF) [9], is making the training of LLMs increasingly expensive and complex.

In facing these challenges, the rapid development of open-source models has shown significant competitiveness. According to the model capability rankings by Chatbot Arena [10], closed-source general LLMs like GPT [6,7], Gemini [11], and Claude [12] rank highly in overall capabilities. However, open-source models are also quickly rising, with notable examples including Llama [13], Qwen [14], Phi [5], and Baichuan [15]. These open-source models each have distinct characteristics and advantages in their architecture and training methods. Llama uses only English data for pre-training, Phi is renowned for its high capability despite a small number of parameters, and Qwen [14], and Baichuan [15] are Chinese-English bilingual models.

Researchers have increasingly focused on developing small-parameter LLMs. Notable examples include Qwen, a compact model featuring a 1.8B parameter version from its inception, as well as recent innovations like Meta's Llama3.2 [16], Google's Gemma2 [17], and Hugging Face's SmolLM2 [18]. These developments underscore a growing emphasis on compact models that balance performance with resource efficiency. Such efforts are critical for addressing scalability challenges while maintaining competitive performance in real-world applications. These advancements, when integrated into existing frameworks, clearly signal a shift towards resource-conscious and specialized AI development.

Despite progress made by general LLMs in handling multilingual tasks, significant challenges remain in specialized fields such as biomedicine [19]. The complex terminology, ambiguous abbreviations, and rare vocabulary in biomedical texts make it difficult for models trained specifically for the English biomedical field to adapt to other language environments [20]. Differences in vocabulary distribution, cultural backgrounds, and lexical mismatches between languages often lead to decreased response quality and generalization ability of the models [21]. Furthermore, the scarcity of high-quality Chinese biomedical data limits the effectiveness of bilingual model pre-training and affects the output quality in various downstream tasks. Therefore, acquiring and efficiently utilizing high-quality biomedical data is key to improving the performance of Chinese-English bilingual models.

Currently, most existing biomedical LLMs have primarily focused on question-answering (QA) tasks. For instance, Med-PaLM2 [22], developed by Google in 2023, achieved high performance on medical QA datasets. Beyond the closed-source models developed by major corporations and research institutions, open-source efforts have also gained traction globally [20,23–25]. Models such as ChatGLM [26], ChatDoctor [27], DoctorGLM [28], HuatuoGPT [23], Ziya-LLaMA [29], and MedChatZH [24] predominantly rely on monolingual (English or Chinese) QA datasets for fine-tuning, with a strong emphasis on improving biomedical QA performance.

Current biomedical LLMs, while excelling in QA tasks, often fail to address broader applications such as cross-lingual medical data processing or biomedical named entity recognition (NER). Efforts like HuaTuoGPT2 [25] and Taiyi [20] have advanced multilingual biomedical modeling, but challenges such as large parameter sizes and limited task adaptability persist. These limitations highlight a pressing need for compact, efficient models that maintain high performance across diverse biomedical tasks.

In the Chinese biomedical field, different base models and training methodologies have been employed to tackle specific tasks, but each approach has limitations. For example, HuaTuoGPT [23], based on Llama 13B, employs complex reinforcement learning methods tailored for Chinese medical QA but lacks scalability. The updated version, HuaTuoGPT2 [25], built on Llama 7B, combines pre-training and fine-tuning strategies but sacrifices its English conversational abilities. MedChatZH [24], based on Baichuan 7B, focuses narrowly on traditional Chinese medicine literature and QA tasks, limiting its versatility across biomedical domains. Taiyi [20], developed on Qwen 7B, employs a bilingual two-stage fine-tuning strategy that enhances its adaptability for biomedical tasks. However, its large model size presents significant obstacles to practical deployment. Current biomedical LLMs, with parameter sizes ranging from 7B to 13B, reveal a pressing research gap: the potential of small-parameter biomedical models to perform effectively across diverse NLP tasks and languages remains largely unexplored. Bridging this gap is essential for overcoming the bottlenecks in both biomedical NLP research and its real-world applications.

To overcome the issues of large parameter sizes in existing models and their inability to comprehensively cover various medical tasks, and to explore the application performance of small-parameter models in the Chinese-English biomedical field, this study developed three versions of BioQwen, small-parameter Chinese-English bilingual biomedical multi-task models, based on Qwen series models. BioQwen adopts a simple and efficient fine-tuning strategy and includes 3 versions with different parameter sizes: BioQwen-0.5B, BioQwen-1.5B and BioQwen-1.8B, designed to handle various biomedical downstream tasks. It also integrates high-quality data preprocessing pipeline to enhance model performance. While maintaining high efficiency, the small-parameter BioQwen significantly reduces computational resource requirements and training costs, with the simplicity of its training method being an additional advantage. The BioQwen models developed in this study, along with their data processing and model training methodologies, not only demonstrates excellent performance in improving text mining capabilities in the biomedical field but also explores the specific performance of small-parameter models in the biomedical field. This promotes the development of Chinese-English bilingual biomedical multi-task models and provides strong support for related research and applications.

The remainder of this paper is organized as follows. Section 2 elaborates on the Materials and Methods, encompassing the research strategy, data collection, preprocessing, and detailed descriptions of the base model, training procedures, and evaluation protocols. Section 3 highlights the key results, including improvements in data cleaning, advancements in domain-specific NER, generative QA tasks, zero-shot English QA performance, and the integration of INT4 quantization for efficient mobile deployment. Section 4 provides a critical discussion of these findings, addressing methodological advantages and limitations. Finally, Section 5 concludes by summarizing the main contributions and proposing directions for future research.

2. Materials and Methods

2.1. Research Strategy

As shown in Fig. 1, we first curated open-source datasets for various biomedical downstream tasks, then conducted data cleaning and formatting, followed by fine-tuning the small-parameter base model, and finally introduced the BioQwen models (three versions). Specifically, for common biomedical downstream tasks such as NER, relation extraction (RE), and machine translation (MT), we collected high-quality Chinese-English bilingual biomedical data. For QA task, we collected high-quality Chinese biomedical data and included Chinese general knowledge data. Then, through a meticulously crafted data preprocessing pipeline and optimized supervised fine-tuning strategies, we achieved a balance of exceptional model performance and efficiency.

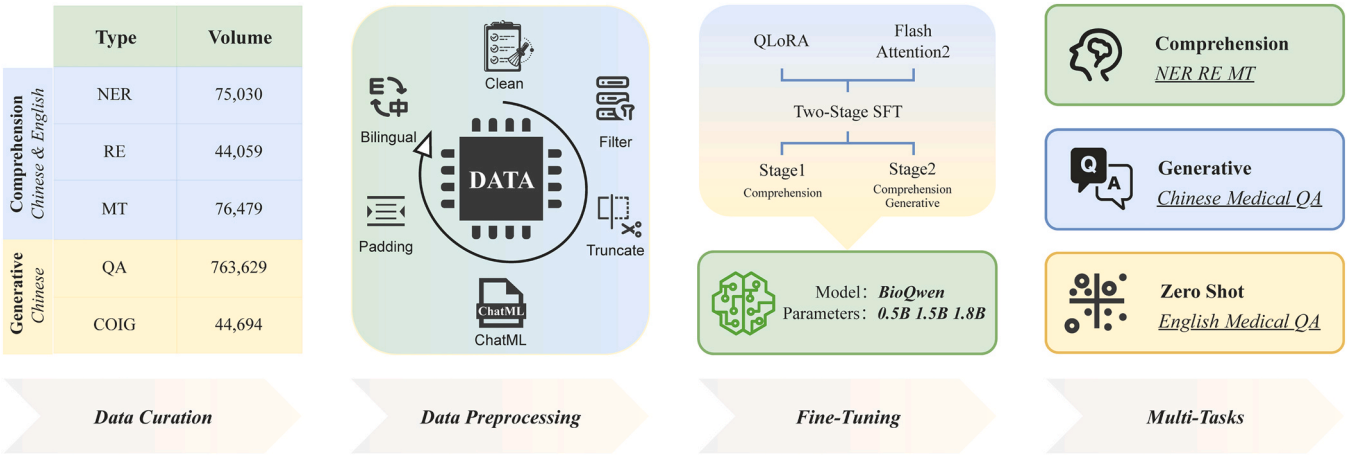

Fig. 1. Workflow for Data Preprocessing and Training of BioQwen Models. The workflow consists of four key stages: (1) Data Curation, involving the curation of both Chinese and English datasets, including biomedical data (NER, RE, MT, QA) and general-domain data (COIG), with the absence of English biomedical QA data; (2) Data Preprocessing, which includes bilingual data cleaning, filtering, truncation, and standardization into the ChatML format, with tailored preprocessing pipeline for bilingual datasets to ensure data quality; (3) Fine-Tuning, employing a two-stage strategy using QLoRA and Flash Attention2 to enhance comprehension and generative capabilities. The first stage focuses on comprehension tasks, while the second integrates generative tasks, resulting in three parameterized versions of BioQwen; and (4) Multi-Task Applications, showcasing the model's versatility across comprehension tasks (NER, RE, MT), generative tasks (Chinese Medical QA), and zero-shot generalization tasks (English Medical QA).

2.2. Data Curation

2.2.1. Comprehension Tasks Data

For comprehension tasks such as NER, RE, and MT, we utilized portions of the open-source datasets fromTaiyi [20]. These datasets have been rigorously reviewed and edited to ensure data accuracy and consistency. The complete collection comprises 140 existing open-source Chinese-English bilingual datasets, including 38 Chinese datasets and 102 English dataset. Specifically, in the NER task, the model needs to identify and classify biomedical entities in the text, such as viruses, diseases, and drugs. The RE task requires the model to extract complex relationships between entities in the text, such as the connection between diseases and symptoms. The MT task involves translating biomedical texts from one language to another, which is particularly important for enhancing the performance of multilingual models.

2.2.2. Generative Tasks Data

The Med-Mix-2M dataset from MedChatZH [24] serves as a foundational resource for our generative tasks, such as QA. This dataset comprises two main components: general-purpose instructions derived from the open-source BELLE-3.5 M [30] dataset and online medical dialog data. The original dataset's general-purpose instructions (1.3 M) and medical dialog data (0.7 M) presented challenges due to their large volume and varying quality. To better accommodate the requirements of training small-parameter models, we replaced the general-purpose instructions with the higher-quality COIG-CQIA [31] dataset (0.03 M), as detailed in Section 2.2.3. Furthermore, we performed extensive cleaning and refinement of the medical dialog data to address quality issues and enhance its overall utility (see Section 2.3).

This dataset provides a rich and diverse collection, primarily sourced from medical health forums, online consultation platforms, and interaction records of medical services within the Chinese internet. By leveraging these authentic dialog samples, the model is trained to communicate naturally and effectively, enabling it to deliver accurate and reliable medical information or advice in Chinese.

2.2.3. Chinese General Knowledge Data

In constructing biomedical LLMs, an over-reliance on highly specialized datasets can hinder the model's performance in general domains [32]. To mitigate this limitation, we specifically introduced the Chinese General Knowledge Dataset COIG-CQIA [31], a systematically curated resource designed to address this challenge. This dataset is derived from a broad array of sources across the Chinese internet, including QA communities, encyclopedia entries, and educational exam materials. These resources cover a wide range of general knowledge topics, providing the model with rich background knowledge and diverse contextual applications.

In addition, the COIG-CQIA research team emphasized the importance of selecting data sources, the necessity of controlling data quality, and the strategy of optimizing model performance by adjusting the data mixing ratio. These contributions offered invaluable guidance for our study, enabling the effective integration and utilization of heterogeneous data sources, thereby enhancing the model's generality and adaptability in addressing non-specialized tasks.

2.3. Data Preprocessing

2.3.1. Data Cleaning

2.3.1.1. Truncate long texts by paragraph. To tackle the issue of the model's difficulty in halting output generation during initial training due to overly long text, we applied a token length limit of 1024 tokens to all data after tokenization. Instead of using a simple truncation method, we retained the entirety of the last complete paragraph to preserve contextual integrity. Specifically, the data was first tokenized to determine the total number of tokens. If the token count exceeded 1024, the text was truncated iteratively by identifying newline positions and removing content paragraph by paragraph until the total token count was less than or equal to 1024. In extreme cases, where no newline characters could be found and the token count still exceeded 1024, it would indicate that the text consisted of a single paragraph with more than 1024 tokens. In such instances, truncation would be performed at the sentence level, using periods as the delimiter. However, practical observations confirmed that this scenario did not occur in the dataset used for this study, validating the robustness of our approach.

The decision to preserve entire paragraphs rather than complete sentences was based on the concept of a paragraph as a cohesive logical unit. Removing clauses or sentences solely based on punctuation (e.g., periods) could compromise the logical integrity of the text. This method preserved textual coherence and was particularly instrumental in training the model to handle interruptions and transitions effectively in real-world conversational scenarios.

2.3.1.2. Remove medical-related content from general data. The COIG-CQIA dataset, included as part of the general data in this study, emphasizes broad applicability and general content. To eliminate any inadvertent inclusion of specialized medical content, the dataset underwent a thorough screening process. For instance, terms such as "medicine," "medical care," or "pharmaceuticals" were systematically excluded through a combination of automated keyword matching and manual review, ensuring the dataset's generality and purity.

2.3.1.3. Integrate reward models and biomedical models for QA data cleaning and visualization. To assess and improve the quality of the raw Chinese QA dataset, we utilized Ziya-LLaMA-7B-Reward [29] to visually analyze the dataset quality. This reward model was crucial for filtering higher-quality biomedical dialog data. The retained data optimized the model's generated dialog content and adhered to professional biomedical communication standards. The refined dataset enabled the generation of responses that were both accurate and naturally expressed, and aligned with ethical standards for biomedical QA. Importantly, the model emphasized that patients should seek timely biomedical attention and follow professional guidance.

To further explore the effectiveness of the data cleaning process, we incorporated two leading biomedical models, BioBERT [33] and BioGPT [34], to compare and visualize data quality before and after cleaning. BioBERT was chosen for its strong performance in biomedical NLP tasks, although its context length limit of 512 tokens required truncation of QA data exceeding this threshold. In contrast, BioGPT was used to handle the complete QA data without truncation, enabling full-length embeddings extraction.

After generation, embeddings from both BioBERT and BioGPT were subjected to Principal Component Analysis (PCA) to reduce dimensions and visualize clustering patterns. By comparing the data distribution before and after cleaning, we evaluated how the reward model improved dataset quality. This combined approach offered insights into dataset refinement and alignment with biomedical communication standards.

2.3.2. Data Formatting

2.3.2.1. Overview. The overall data normalization process is shown in Fig. 2. The data cleaned in the previous step are converted into a specific format, including example dialogs from the system, user, BioQwen, and other components. These dialog samples use tagged syntax to indicate the start and end of different dialog participants and content. Finally, these normalized data are input into the training process of BioQwen, enabling the model to enhance its comprehension and dialog capabilities in the biomedical field through learning this structured dialog data.

2.3.2.2. ChatML Format. In this study, all data were converted to the ChatML [14] format, a format proposed by OpenAI specifically designed for dialog systems to clearly define dialog flow. This format encapsulates each part of the dialog using specific tags, making model training and dialog management more efficient.

2.3.2.3. Special Padding Tokens. To address the issue of varying data lengths causing the model to inaccurately stop output during data processing, we used special padding tokens instead of just traditional end tokens. These padding tokens in the ChatML format ensure that all dialogs have consistent length before being fed into the model, without disrupting the natural flow and semantic content of the dialogs. This method effectively resolves the problem of the model potentially producing repetitive output when handling long texts [35].

2.3.2.4. Multilingual Adaptation. When processing Chinese-English bilingual datasets, each dialog is configured with system prompts corresponding to its language characteristics. To further enhance the model's performance in a multilingual environment, we specifically included a Chinese-English translation (MT) task during the model's fine-tuning stage. This is intended to enable the model to learn direct translation skills between languages, preventing it from losing its original English dialog capability and helping the model understand and internalize the subtle differences in expressing the same concepts in both languages.

2.4. Base Model

The Qwen series of open-source models has provided significant convenience and support for our research. These pre-trained models possess strong multilingual processing capabilities. By introducing specific language embeddings and enhanced self-attention mechanisms, they effectively recognize and optimize translation and understanding between different languages, supporting research on multilingual

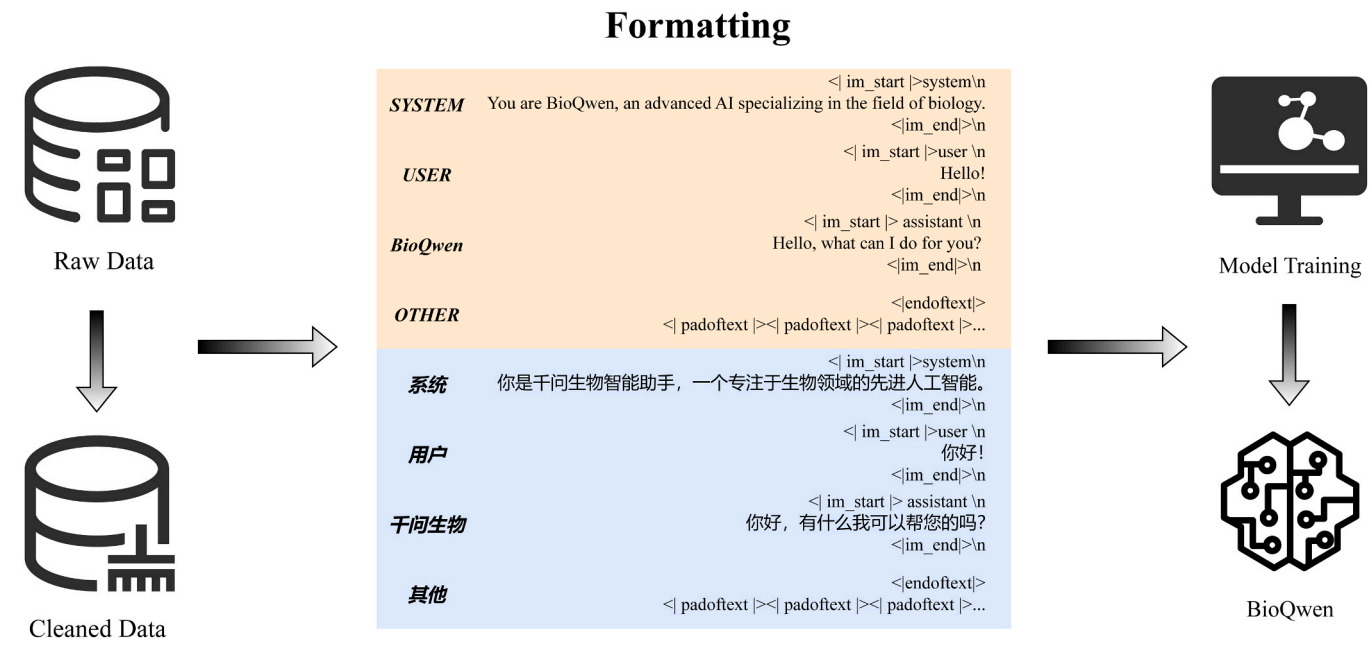

Fig. 2. Schematic diagram of the data formatting process.

datasets [14]. Additionally, the Qwen models offer multiple parameter versions to meet different research needs. Their outstanding performance and strong language comprehension capabilities in the latest public benchmarks [36] laid a solid foundation for our downstream tasks. For our experiments, we selected the smallest parameter versions from the Qwen series: Qwen1.5 (1.8B), Qwen2 (0.5B), and Qwen2.5 (1.5B). After training, we renamed them as BioQwen-0.5B, BioQwen-1.5B, and BioQwen-1.8B based on their parameter sizes, to comprehensively explore the performance of small-parameter models in the Chinese-English biomedical field.

2.5. Model Training

During the model training process, we adopted optimized fine-tuning techniques such as QLoRA [37] and FlashAttention-2 [38], while implementing a two-stage fine-tuning strategy inspired by prior works [20,39]. In our study, biomedical downstream tasks were classified into comprehension tasks and generative tasks. In the first stage, the model was trained on comprehension tasks such as NER. In the second stage, data from generative tasks such as QA were added, and the model was trained with a combination of both comprehension and generative task data.

In terms of hyperparameter settings, these values were determined through a combination of empirical observations, experimental validation, and insights from the literature [20,39]. For instance, the learning rate for the first stage is set to 2e-4, using the AdamW optimizer to adjust weights. Smaller learning rates (1e-6 and 1e-5) resulted in slow convergence, while larger rates (3e-4) caused significant fluctuations in the loss curve. The value 2e-4 provided an optimal balance, achieved a balance of stability and efficiency. A 10 % warm-up period is employed, during which the learning rate gradually increases with a linear scheduler to mitigate early training instability. The batch size per computing device is set to 4, combined with a gradient accumulation strategy of 4 steps. In the Lora configuration, the r value is set to 64, alpha is set to 16, and dropout is set to 0.05 to prevent overfitting. The second stage of training focuses on adapting the model to complex biomedical dialog data, maintaining a learning rate of 2e-4 while introducing a small weight decay of 1e-5. The warm-up period remains at 10 % of the total training steps, and the training cycle is extended to 5 epochs to ensure the model fully learns and adapts to more complex unstructured content. The number of epochs was determined by monitoring the loss during training, with a target loss below 1 serving as an early indicator of appropriate fitting. This threshold was validated through subsequent evaluation processes to confirm that the model learned and adapted to unstructured content.

Using four 3090 (24 G) GPUs, BioQwen-0.5B required 5 hours for the first stage and 18 hours for the second stage. For BioQwen-1.5B, training required 6 hours for the first stage and 39 hours for the second stage, a notable increase compared to BioQwen-0.5B. Similarly, BioQwen-1.8B required 7 hours for the first stage and 44 hours for the second stage, reflecting a continued increase in training time.

2.6. Model Evaluation

To evaluate the model's performance on the NER task, we employed *per − class F*1, *macro F*1, and *balanced accuracy* as the primary evaluation metrics. These metrics were selected to ensure a comprehensive assessment of the model's performance across all classes, particularly in biomedical entity recognition tasks.

Specifically, we measured *Chemical F*1, *Disease F*1 and *Drug F*1 to capture the distinct performance of the model on each entity type, reflecting how well the model performs on each category individually. We additionally calculated the *Macro F*1 score to ensure equal weighting of performance across classes, mitigating the influence of class imbalance, and *Balanced Accuracy* to address potential class imbalance issues. Formally, these metrics can be written as follows:

$$F_1^{(\text{Chemical})} = 2 \times \frac{\text{Precision}_{\text{Chemical}} \times \text{Recall}_{\text{Chemical}}}{\text{Precision}_{\text{Chemical}} + \text{Recall}_{\text{Chemical}}} \#\# \tag{1}$$

$$F_1^{(\text{Disease})} = 2 \times \frac{\text{Precision}_{\text{Disease}} \times \text{Recall}_{\text{Disease}}}{\text{Precision}_{\text{Disease}} + \text{Recall}_{\text{Disease}}} \tag{2}$$

$$F_1^{(\text{Drug})} = 2 \times \frac{\text{Precision}_{\text{Drug}} \times \text{Recall}_{\text{Drug}}}{\text{Precision}_{\text{Drug}} + \text{Recall}_{\text{Drug}}} \tag{3}$$

$$F_1^{(\text{Macro})} = \frac{1}{C}\sum_{i=1}^{C} F_1^{(i)} \# \tag{4}$$

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{5}$$

where Precision_X and Recall_X are the precision and recall for the "X" class (e.g., X=*Chemical*). $F_1^{(i)}$ is the $F1$ score for the i-th class, C is the total number of classes (e.g., C=2 for *Chemical* and *Disease*), and TP_i, FN_i denote the true positives and false negatives for the i-th class, respectively.

Following prior work [19,40,41], we used three widely-adopted benchmarks to evaluate our proposed approach: NCBI Disease Corpus (NCBI, en) [42], BioCreative V CDR Task Corpus(BC5CDR, en) [43], and CMeEE (zh) [44]. These corpora cover both English and Chinese biomedical text, allowing us to assess the generalizability of our model across languages and various biomedical entity types.

To comprehensively evaluate the model's performance on biomedical QA tasks, we used BLEU [45], GLEU [46], and ROUGE [47] metrics to measure the quality of the generated text. These internationally recognized metrics are used to evaluate the accuracy, fluency, and coverage of the text. Specifically, they measure different aspects of the models' performance, including the precision of n-grams (BLEU scores), recall of unigrams and bigrams (ROUGE scores), the longest common subsequence (ROUGE L), and a balanced measure of precision and recall (GLEU).

Additionally, to demonstrate the model's efficiency, we introduced Iteration Time as an evaluation metric. This not only reflects the model's task processing speed but also provides an important performance baseline for deploying the model on mobile devices. We selected two Chinese datasets, WebMedQA [48] (zh) and cMedQA2 [49] (zh), for evaluation, and used iCliniq [27] (en) English biomedical dialog data to assess its zero-shot learning capability.

3. Results

3.1. Model-based data cleaning improves clustering accuracy and ensures reliability

This study adopted a strategy based on reward model scoring to thoroughly clean the raw Chinese QA data. The quality score distribution of the entire QA dataset is shown in Fig. 3a, where it can be intuitively observed that some data fall into the low-quality region, necessitating further cleaning. After analyzing the score distribution of the entire dataset, a threshold value of 2.80 was selected for dataset cleaning. This threshold was chosen to maintain consistency in data volume with comprehension tasks, which aids in improving the stability of model training [31]. The threshold is based on the distribution characteristics of data quality distribution in the chart and is used to remove data below this value, ensuring the high quality of the remaining data. The quality score distribution of the cleaned data is illustrated in Fig. 3b.

Regarding the clustering patterns, Fig. 3c and Fig. 3d illustrate PCA-based dimensionality reduction results applied to embeddings from BioBERT and BioGPT, respectively. In Fig. 3c, BioBERT embeddings, constrained by the model's input limit of 512 tokens, were used. As a result, the PCA clustering exhibits limited resolution, with the data

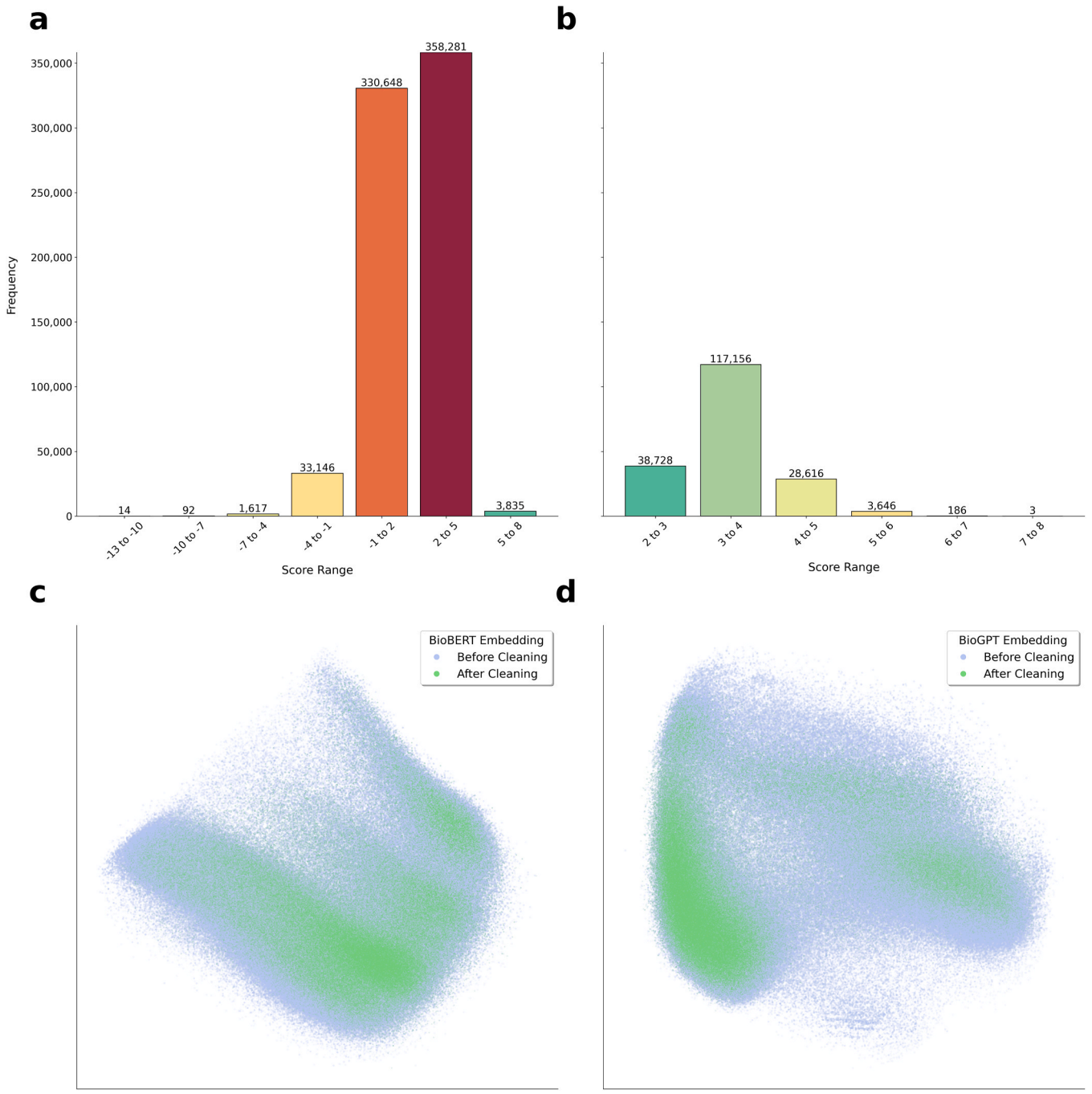

Fig. 3. Distribution of quality scores before and after cleaning the biomedical QA data.

roughly grouped into three clusters. Notably, the cleaned data spans all three clusters, with the majority concentrated in the largest cluster. Additionally, most outliers were effectively removed during the cleaning process.

In Fig. 3d, BioGPT embeddings were utilized, enabling complete input data processing without the token limit constraints associated with BioBERT. Consequently, the PCA clustering shows greater separation between clusters, with the data primarily grouped into two clusters. After cleaning, the majority of data concentrates in the largest cluster on the plot's left side. A smaller portion is distributed in the cluster on the right side, while a considerable number of outliers have been removed. This demonstrates that the cleaning process enhanced the dataset quality and clustering patterns.

The comparison of the data distribution and tokens before and after cleaning the entire dataset is shown in Table 1. From the charts, it can be observed that although the data volume for each task type generally

Table 1
Data Distribution and Total Tokens Used for Each Task Type Before and After Data Cleaning.

Task Type	Raw Volume	Cleaned Volume	Data Ratio	Number of Tokens (M)
NER	75,030	66,932	16.7 %	14.5
RE	44,059	42,089	10.8 %	9.3
MT	76,479	73,748	9.7 %	8.5
QA	763,629	188,335	48.8 %	42.5
COIG	44,694	33,003	14.0 %	12.2
Total	**1003,891**	**404,107**	**100 %**	**87.0**

decreased, the extent of reduction varies. Specifically, the most significant reduction occurred in the biomedical QA task, mainly due to the removal of a large amount of duplicate or low-quality data. Additionally, the data volumes for NER and RE tasks also decreased but to a lesser extent, indicating that the raw data quality for these tasks was relatively high.

3.2. BioQwen achieves superior domain-specific NER performance via two-stage fine-tuning

Given that most biomedical models focus on biomedical QA tasks, such as MedChatZH and HuaTuoGPT2, currently, only the Taiyi model includes data from NER tasks during the training phase. GPT-3.5-turbo, being the most popular free closed-source model, is also used as a benchmark for comparison, with specific data sources referenced from the aforementioned review articles. Detailed evaluation results are shown in Table 2.

Table 2 summarizes the performance of various models on three NER datasets: BC5CDR (en), NCBI (en), and CMeEE (zh). The evaluation includes per-class F1 scores (e.g., Chemical F1, Disease F1, Drug F1), overall Macro F1 and Balanced Accuracy metrics. Bolded values highlight the best results for individual metrics. The table compares BioQwen models across two fine-tuning stages (Stage 1 and Stage 2) with the Taiyi (7B) model and the general-purpose GPT3.5 Turbo (175B) model. The results in Table 2 reveal that two-stage fine-tuning (Stage 2) significantly improves task performance. For instance, BioQwen-1.8B Stage 2 achieves the highest Macro F1 score (0.730) on BC5CDR and the highest Balanced Accuracy (0.808) on CMeEE. However, BioQwen-1.5B in both Stage 1 and Stage 2 underperforms compared to the smaller BioQwen-0.5B Stage 2, particularly on BC5CDR and NCBI. This unexpected result is likely attributable to the Qwen2.5 official base model, which incorporates multilingual enhancements. Since our training data comprises only Chinese and English bilingual content, this design may introduce irrelevant outputs, thereby degrading performance on NER tasks. A more detailed discussion regarding the performance of BioQwen-1.5B will be provided in Section 4.

Smaller and more focused models, such as BioQwen-0.5B Stage 2, exhibit stability and consistency. This highlights the importance of targeted fine-tuning for domain-specific tasks, particularly when sufficient bilingual training data is available. These results illustrate the trade-offs between larger and smaller models: larger multilingual models offer greater capacity but require extensive, high-quality domain-specific data to realize their potential in specialized domains like biomedical NER.

BioQwen-1.8B and BioQwen-0.5B both demonstrate strong performance after two-stage fine-tuning, with notable improvements observed across all evaluation datasets. However, while the Taiyi model performs well on the CMeEE-Disease (zh) dataset, its overall performance is slightly inferior to BioQwen. While lagging behind in these tasks, GPT-3.5-turbo offers valuable reference as a free, closed-source model.

3.3. BioQwen delivers efficient generative QA with fewer parameters

Fig. 4 illustrates the parameter relationships between open-source biomedical LLMs with conversational capabilities and small-parameter LLMs. The figure compares parameter counts (in billions) across two categories: (a) biomedical LLMs and (b) small-parameter LLMs.

In the biomedical LLMs category (Fig. 4a), HuatuoGPT [23] and Ziya-LLaMA [29] have the largest parameter counts at 13B. They are followed by HuaTuoGPT2, MedChatZH, and Taiyi, which are each configured with 7B parameters. DoctorGLM [28] and ChatGLM [26] have 6B parameters. In contrast, our BioQwen models prioritize parameter efficiency, with BioQwen-1.8B, BioQwen-1.5B, and BioQwen-0.5B. Studies indicate that models like ChatGLM, DoctorGLM, HuatuoGPT, and Ziya-LLaMA perform relatively weaker on biomedical tasks, leading this study to focus on Taiyi, MedChatZH, and HuaTuoGPT2.

In the small LLM category (Fig. 4b), additional baselines were added for robust comparisons with our BioQwen models. Specifically, Qwen2.5-Instruct (1.5B) and Qwen1.5-Chat (1.8B) were included as reference models to compare with our BioQwen-1.5B and BioQwen-1.8B. These Qwen models, trained through SFT and RLHF, serve as robust baselines owing to their conversational capabilities and task adaptability. Since base models lack dialog capabilities, these officially fine-tuned models were used for comparisons, ensuring fairness in evaluation. Other models in this category include Gemma2 (2B), SmolLM2 (1.7B), and Llama3.2 (1B). Notably, BioQwen-0.5B is the smallest, with only 0.5B parameters.

The performance of each model on the WebMedQA (zh) dataset is shown in Fig. 5a (detailed in Supplementary Table 1), and their performance on the cMedQA2 (zh) dataset is shown in Fig. 5b (detailed in Supplementary Table 2). The figure demonstrate that the BioQwen models, despite their smaller parameter sizes, outperform larger models in multiple metrics, highlighting their efficiency and effectiveness. In comparison, the Taiyi, MedChatZH, and HuaTuoGPT2 models have parameter sizes of 7B, while the Qwen1.5-Chat model was trained using complex methods such as RLHF. The comparison shows that BioQwen models, while maintaining smaller parameter sizes, still exhibit strong performance and efficiency, which is significant for biomedical QA systems that need to be deployed in resource-constrained environments.

To evaluate the runtime efficiency of different models on biomedical QA tasks, we conducted tests on the aforementioned two Chinese datasets, using the average time per iteration (seconds) as the metric. The evaluation results are shown in Fig. 5c. The results indicate that due

Table 2
NER Task Performance Comparison Across Different Models and Fine-Tuning Stages.

Model	Parameters	BC5CDR (en)				NCBI (en)		CMeEE (zh)			
		Chemical F1	Disease F1	Macro F1	Balanced Accuracy	Disease F1	Accuracy	Disease F1	Drug F1	Macro F1	Balanced Accuracy
BioQwen−0.5B Stage1	0.5B	0.789	0.599	0.694	0.773	0.677	0.753	0.656	0.764	0.710	0.742
BioQwen−0.5B Stage2	0.5B	0.809	0.628	0.719	0.784	0.717	**0.778**	0.705	0.799	0.752	0.773
BioQwen−1.5B Stage1	1.5B	0.779	0.454	0.616	**0.818**	0.582	0.693	0.581	0.631	0.606	0.636
BioQwen−1.5B Stage2	1.5B	0.781	0.588	0.684	0.793	0.606	0.707	0.676	0.731	0.704	0.676
BioQwen−1.8B Stage1	1.8B	0.676	0.521	0.599	0.731	0.516	0.664	0.675	0.807	0.741	0.755
BioQwen−1.8B Stage2	1.8B	**0.819**	**0.640**	**0.730**	0.802	0.694	0.767	0.741	**0.853**	**0.797**	**0.808**
Taiyi	7B	0.741	0.629	0.685	0.757	**0.724**	0.776	**0.823**	0.750	0.787	0.803
GPT3.5 turbo	175B	0.603	0.518	-	-	0.505	-	0.470	0.470	-	-

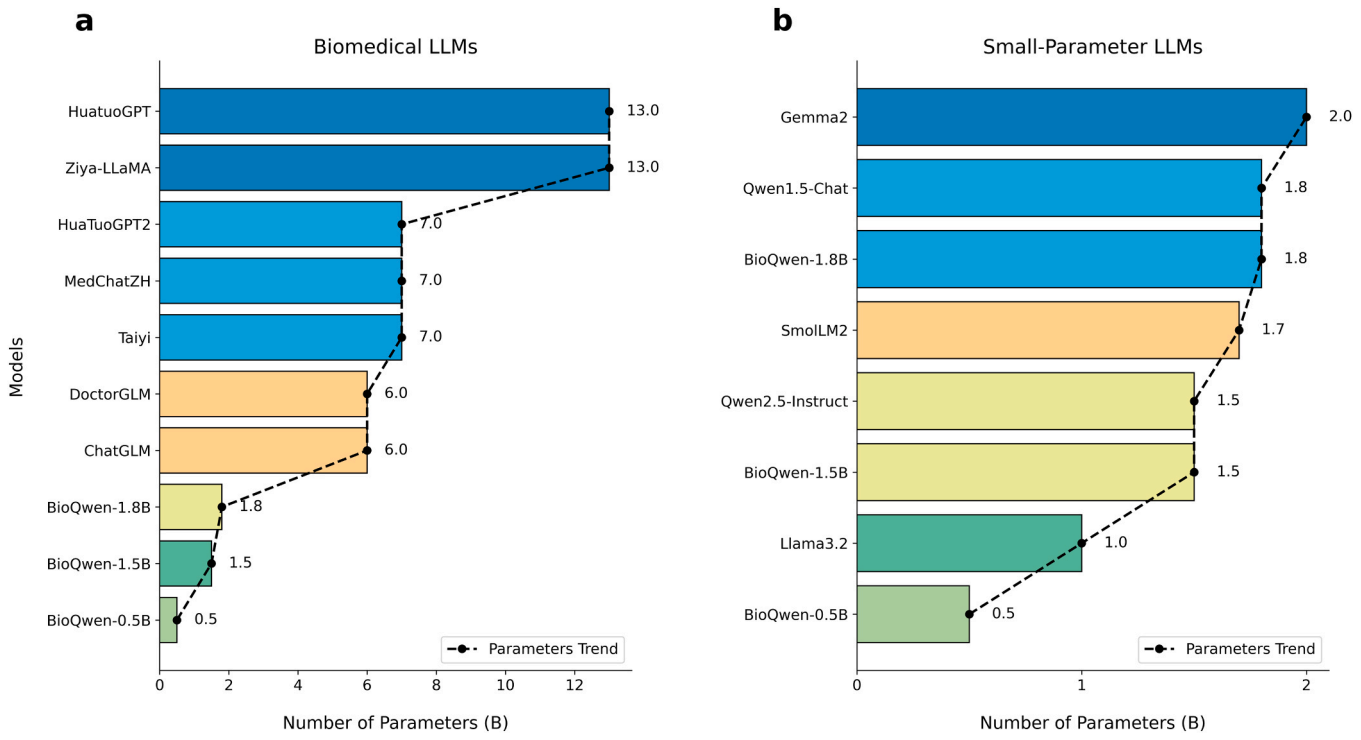

Fig. 4. Bar chart illustrating the parameter sizes of biomedical LLMs, small-parameter LLMs, and BioQwen. (a) Depicts the parameter sizes of biomedical LLMs, with HuatuoGPT and Ziya-LLAMA leading at 13 billion parameters, followed by HuaTuoGPT2, MedChatZH and Taiyi at 7 billion. (b) Shows the parameter sizes of small-parameter LLMs, with Gemma2 leading at 2 billion parameters. The dashed line represents the trend of parameter sizes within each group.

to their smaller parameter sizes, the iteration times for BioQwen-0.5B, BioQwen-1.5B and BioQwen-1.8B on the two datasets were 4.09 seconds, 3.88 seconds and 3.42 seconds, as well as 3.92 seconds,3.90 seconds and 3.79 seconds, respectively, significantly lower than those of larger parameter models such as Taiyi (10.86 seconds and 14.16 seconds) and HuaTuoGPT2 (13.45 seconds and 13.11 seconds). These results demonstrate that BioQwen maintains high performance while achieving higher runtime efficiency, highlighting their great potential in resource-constrained environments.

Interestingly, despite having only 2 billion parameters, Gemma2 exhibited relatively slow iteration times of 16.97 seconds and 15.86 seconds across the two datasets. Further analysis revealed that Gemma2's outputs often contained excessive non-biomedical content, leading to low evaluation scores in biomedical QA tasks. This issue may result from the official training strategies employed for Gemma2, including RLHF and similar approaches, which appear to bias the model toward generating verbose outputs. However, in the context of biomedical QA, verbosity does not equate to quality; longer responses are often unnecessary and may dilute relevance. These findings highlight the critical role of domain-specific fine-tuning in ensuring that models remain efficient and relevant for specialized tasks.

3.4. BioQwen excels in zero-shot English QA despite being monolingually fine-tuned

To evaluate the zero-shot learning capabilities of BioQwen on English biomedical dialog tasks, we collected real doctor-patient dialog data from the iCliniq website for testing. Notably, BioQwen-0.5B was excluded from this evaluation due to its inability to consistently maintain English conversational capabilities, occasionally generating responses in Chinese. A similar phenomenon was observed in other small-parameter models, such as Llama3.2, which sometimes responded in English when tested on Chinese datasets like WebMedQA and cMedQA2. This inconsistency highlights the challenges small-parameter models face in maintaining stable bilingual dialog abilities. MedChatZH and HuaTuoGPT2 were excluded from the evaluation as they completely lack English conversational capabilities.

Despite being fine-tuned exclusively on Chinese biomedical QA data and having no exposure to English QA data during training, BioQwen-1.5B and BioQwen-1.8B were evaluated against other biomedical LLMs and small-parameter LLMs capable of bilingual (Chinese-English) dialog. The compared models include Taiyi, Qwen1.5-Chat, Qwen2.5-Instruct, and 3 smaller-parameter models such as Llama3.2 and Gemma2.

As shown in Fig. 5d (detailed in Supplementary Table 3), BioQwen-1.8B outperformed Taiyi, Qwen1.5-Chat, Qwen2.5-Instruct, and small-parameter models across all metrics, including BLEU, GLEU, and ROUGE, with particularly significant advantages in BLEU 3, BLEU 4 and ROUGE 2 scores. This superior performance can likely be attributed to the increased parameter size of BioQwen-1.8B, which enhances its capacity to generalize and adapt to unseen tasks, such as English biomedical dialogs. Larger parameter models generally benefit from richer representational power, allowing them to better capture the complexities of diverse languages and contexts.

Interestingly, all other models, such as Taiyi, were trained using English QA data, their performance on these metrics was slightly inferior, further underscoring BioQwen-1.8B's strong zero-shot learning ability. This indicates that, despite being trained solely on Chinese QA data, BioQwen-1.8B can generate high-quality responses in English, demonstrating remarkable adaptability. Its exceptional performance not only validates its robust bilingual capabilities but also highlights its potential for broader applications in bilingual Chinese-English environments.

3.5. BioQwen enables efficient mobile deployment via INT4 quantization

Leveraging the MLC-LLM [50] project, we successfully deployed the BioQwen onto mobile devices. This was achieved by streamlining the development and deployment process via the MLC-LLM Android SDK. By integrating INT4 model quantization, a method of reducing model

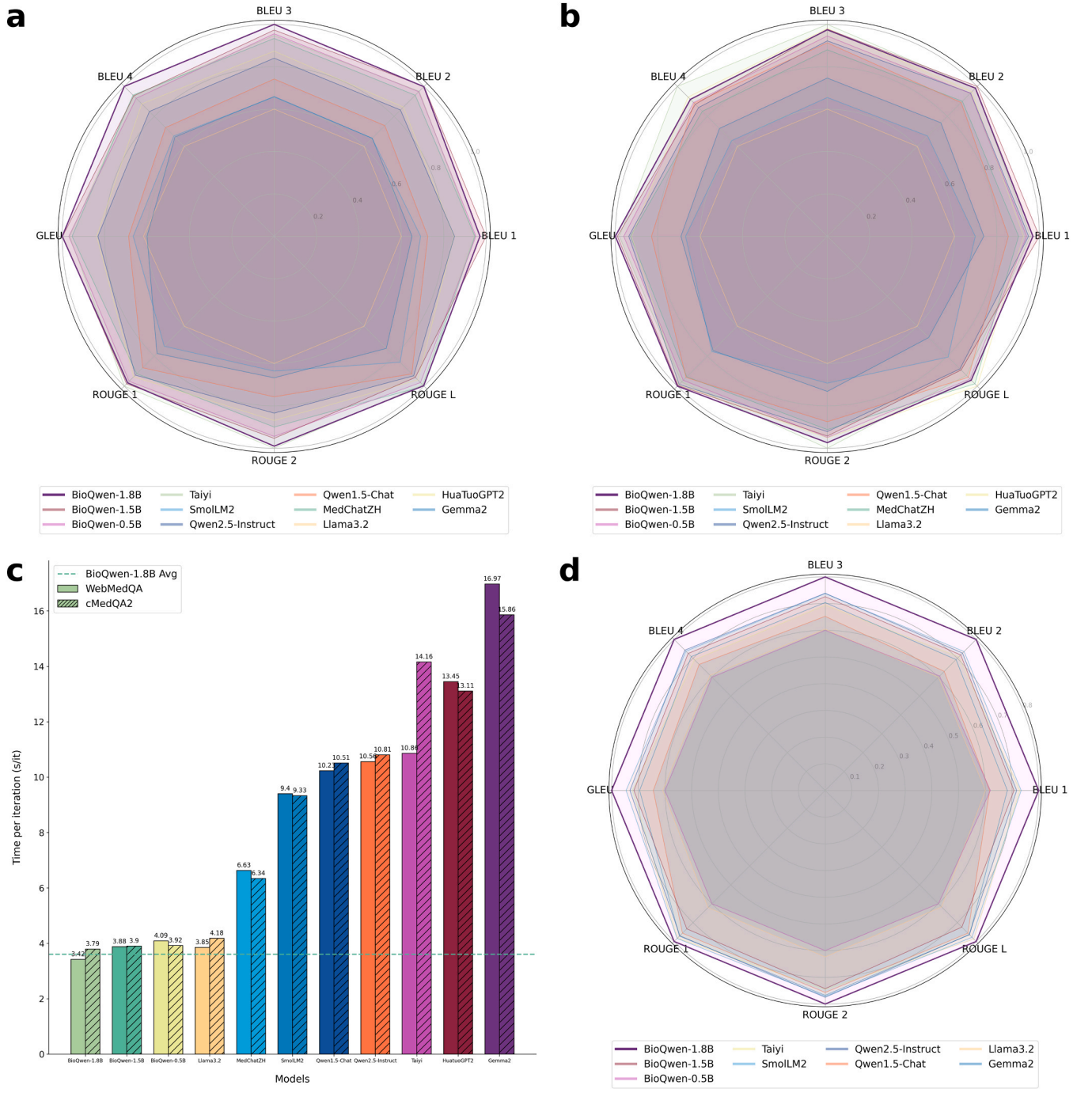

Fig. 5. Performance and operational efficiency comparison of various models across multiple datasets. (a) Radar chart illustrating the performance metrics of various models on the WebMedQA dataset. (b) Radar chart illustrating the performance metrics of various models on the cMedQA2 dataset. (c) Comparative analysis of the operational efficiency of various models on the WebMedQA and cMedQA2 datasets. (d) Radar chart illustrating the performance metrics of various models on the iCliniq dataset. For a detailed descriptions of the evaluation metrics for each axis can be found in the Section 2.6.

size and computational complexity by representing weights and activations with 4-bit integers, we reduced the size of BioQwen-1.5B to under 800 MB and BioQwen-1.8B to under 1 GB, ensuring their compatibility with mobile device constraints. The deployment process involved building runtime libraries, customizing model weight configurations, and utilizing GPU-accelerated execution for efficiency. Given BioQwen's slightly lower performance on cMedQA2 compared to WebMedQA, we selected cMedQA2 for mobile testing to test its applicability for challenging biomedical queries. The results are presented in Supplementary Figs. 1 and 2.

To ensure the efficiency and practicality of deploying INT4-quantized models on mobile devices, we first conducted a detailed evaluation of INT4 quantization's performance impact. Using the evaluation procedures described in Sections 3.3 and 3.4, we assessed the performance impact of INT4 quantization on BioQwen-1.5B and BioQwen-1.8B across three QA datasets. Table 3 presents the results for cMedQA2, while Supplementary Tables 4 and 5 present the evaluations for WebMedQA and iCliniq. Quantization introduced minor performance degradation across all metrics on cMedQA2, aligning with observations from other datasets. For example, BioQwen-1.8B INT4 retains a BLEU-1 score of 17.603 and a ROUGE-L score of 18.108, while maintaining competitive performance and significantly reducing

Table 3
QA Task Performance Comparison and Quantization Analysis of BioQwen Models on the cMedQA2 Dataset.

Model	Precision	Total Size (GB)	BLEU −1	BLEU −2	BLEU −3	BLEU −4	GLEU	ROUGE −1	ROUGE −2	ROUGE -L
Qwen2.5-Instruct 1.5B	bfloat16	3	17.181	8.990	4.577	2.317	6.416	30.564	7.654	16.649
Qwen1.5-Chat 1.8B	bfloat16	3.6	14.247	7.914	4.487	2.488	5.290	28.730	7.003	17.890
BioQwen−0.5B	bfloat16	1	16.899	8.963	4.800	2.512	6.666	29.466	8.040	18.148
BioQwen−1.5B INT4	int4	0.75	18.323	9.461	4.967	2.321	6.629	29.960	7.812	16.648
BioQwen−1.5B	bfloat16	3	**18.738**	**9.744**	**5.117**	2.444	6.981	30.845	7.949	16.822
BioQwen−1.8B INT4	int4	0.9	17.603	9.082	4.843	2.500	6.807	30.350	8.004	18.108
BioQwen−1.8B	bfloat16	3.6	17.811	9.545	5.084	**2.633**	**7.083**	**31.013**	**8.389**	**18.315**

computational overhead.

Table 3 summarizes the QA task performance on the cMedQA2 dataset for BioQwen-1.5B and BioQwen-1.8B before and after INT4 quantization. The results are benchmarked against their official counterparts, Qwen2.5-Instruct (1.5B) and Qwen1.5-Chat (1.8B), as well as the unquantized BioQwen-0.5B. Notably, BioQwen outperforms their official counterparts, with BioQwen-1.5B achieving a BLEU-1 score of 18.738 and a BLEU-2 score of 9.744, surpassing the scores of Qwen2.5-Instruct (BLEU-1: 17.181; BLEU-2: 8.990). Similarly, BioQwen-1.8B exhibits superior performance, further validating the advantages of these models. The findings underscore the effectiveness of domain-specific fine-tuning in enhancing task performance.

Moreover, quantization using INT4 significantly reduces model size while maintaining competitive performance. For instance, BioQwen-1.5B INT4 achieves a BLEU-1 score of 18.323 and a ROUGE-L score of 16.648, closely approximating the scores of its unquantized counterpart (BLEU-1: 18.738; ROUGE-L: 16.822). Likewise, BioQwen-1.8B INT4 demonstrates a BLEU-1 score of 17.603 and a ROUGE-L score of 18.108, indicating minimal performance degradation despite the substantial reduction in model size. These results highlight the practicality of lightweight, quantized models for deployment in resource-constrained environments.

Interestingly, the quantized versions of BioQwen-1.5B and BioQwen-1.8B deliver performance levels comparable to the unquantized BioQwen-0.5B, with varying outcomes across metrics. However, as discussed in Section 2.4, BioQwen-0.5B struggles to consistently sustain English conversational capabilities, revealing the limitations of smaller parameter models. In contrast, the larger parameter models, BioQwen-1.5B and BioQwen-1.8B, exhibit stronger capabilities, including enhanced zero-shot performance. These observations may suggest that, under the same training conditions, leveraging larger parameter models with quantization is a more effective strategy than relying solely on smaller parameter models for deployment.

Based on these evaluations, we selected the quantized BioQwen-1.8B INT4 model, balancing strong performance and compact size, for deployment on mobile devices to evaluate its applicability in real-world scenarios. To rigorously test its performance in real-world biomedical tasks, we conducted mobile deployment tests using the cMedQA2 (zh) dataset. This dataset, known for its challenging nature, provided a robust benchmark for evaluating the model's capacity to handle complex biomedical queries. Supplementary Figs. 1 and 2 illustrate test results for randomly selected questions on diverse topics, including HIV, urticaria, and heart disease during pregnancy. The responses are presented in their original Chinese and corresponding English translations. The results demonstrate that the models produce accurate and empathetic biomedical responses, highlighting their reliability and utility in medical contexts.

In conclusion, these findings confirm the benefits of domain-specific fine-tuning and quantization for improving performance and efficiency in biomedical QA tasks. The quantized BioQwen-1.5B and BioQwen-1.8B models achieve strong performance with significantly lower resource requirements, enabling successful deployment on mobile devices. This demonstrates the feasibility of deploying advanced biomedical QA systems in resource-constrained real-world environments.

4. Discussion

Our evaluation of comprehension tasks (summarized in Table 2 and visualized in Fig. 6a) generally indicated that BioQwen-1.8B, the model with the largest parameter size, achieved the highest performance, aligning with prior research suggesting a positive correlation between model parameter size and performance under consistent training data and methods [4]. However, it's important to note that performance differences were not uniformly substantial across all datasets, with some instances of closely matched results. For example, while BioQwen-1.8B demonstrated the best performance on BC5CDR-Chem (en) with a score of 0.819, the scores for BioQwen-0.5B (0.809) and BioQwen-1.5B (0.781) were remarkably competitive, highlighting nuanced performance variations.

Contrary to our initial expectation of it serving as a performance midpoint, BioQwen-1.5B demonstrably underperformed in comprehension tasks. Across all measured comprehension metrics, its performance was consistently the poorest, even falling below that of BioQwen-0.5B. While the observed language mixing issue (evidenced by the output of unrelated languages like Thai during NER evaluations, leading to imperfect entity matching) likely contributed to its subpar performance in NER tasks, the consistently low scores across the broader spectrum of comprehension metrics suggest a more fundamental underlying issue. The fine-tuning of BioQwen-1.5B, based on the multilingual and long-text focus of the Qwen2.5 pre-trained model [51], appears not to have translated effectively into improved performance on our specific comprehension datasets, which are limited to English and Chinese. Furthermore, the relatively small proportion of NER data within our training data (16.7 %, as detailed in Table 1) diminishes the likelihood that poor NER performance alone is the sole driver of its overall weak comprehension abilities.

In stark contrast to its comprehension performance, BioQwen-1.5B performed as anticipated in generative tasks (detailed in Fig. 5 and Supplementary Tables 1, 2, and 3). Here, it exhibited a clear advantage over BioQwen-0.5B and, in some instances, approached or even slightly surpassed the performance of BioQwen-1.8B. This discrepancy suggests that the larger QA dataset volume inherent in generative tasks was better aligned with the pre-training emphasis of Qwen2.5 on multilingualism and longer sequence processing. This divergence in performance underscores the critical importance of aligning the characteristics of fine-tuning data with the underlying pre-training paradigm [39].

Subsequent to the second stage of learning (as depicted in Fig. 6b), we observed a differential impact on model performance based on parameter size. BioQwen-0.5B, with its smaller parameter count, demonstrated less improvement compared to the larger BioQwen-1.8B. Specifically, the improvement in BioQwen-1.8B's performance was more pronounced on English datasets, while gains on Chinese datasets were less evident. Intriguingly, BioQwen-0.5B exhibited greater improvement than BioQwen-1.8B on the CMeEE-Drug (zh) dataset. This observation lends support to the hypothesis that the parameter size of the base model dictates its performance ceiling. Models with larger parameter capacities possess a higher potential performance ceiling, whereas smaller parameter models may reach their performance limits more rapidly. Following the initial stage of training, it appears that BioQwen-0.5B had largely reached its performance plateau for

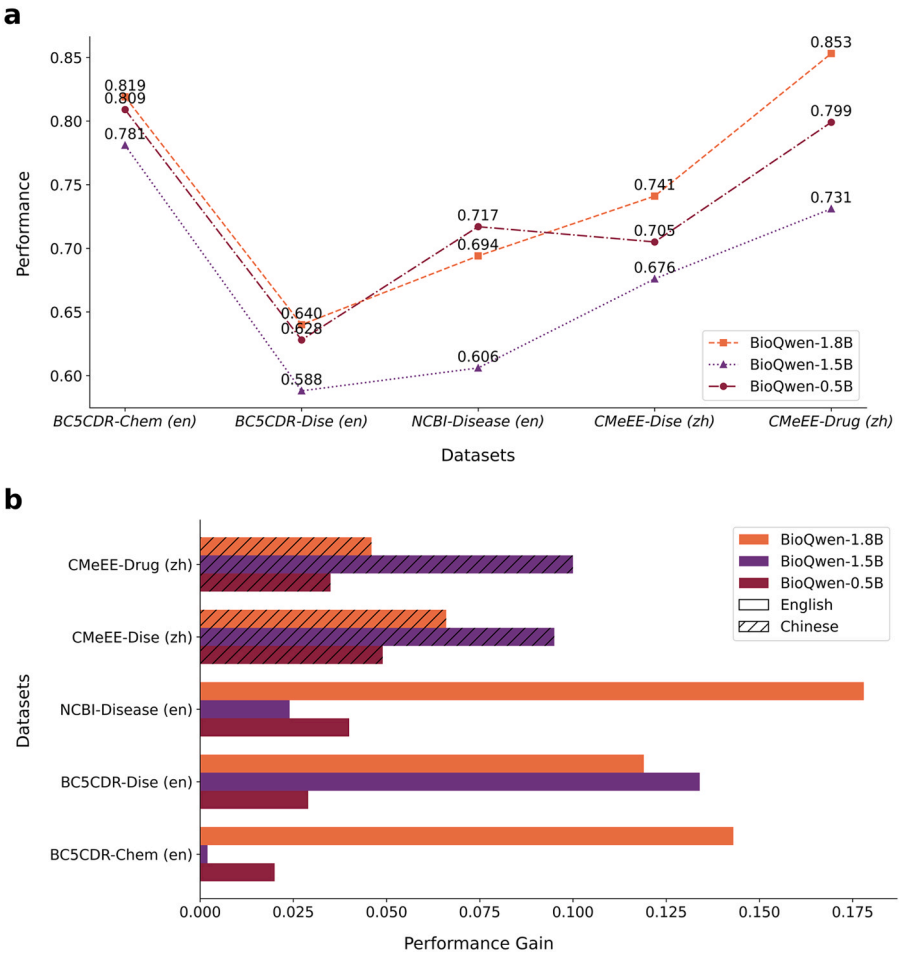

Fig. 6. Performance comparison of different parameter versions of BioQwen in various tasks. (a) Performance comparison of different parameter versions of BioQwen in NER tasks. (b) Performance improvement comparison of different parameter versions of BioQwen in comprehension tasks after the second stage of training.

comprehension tasks, resulting in diminished gains during the subsequent stage. Conversely, BioQwen-1.8B retained significant potential for improvement after the initial training phase.

Regarding generative tasks, BioQwen-1.8B consistently outperformed BioQwen-0.5B, which is likely attributed to the nature of unstructured data prevalent in these tasks, necessitating larger models for effective data fitting. The increased capacity of larger models allows for better learning of the complexities inherent in unstructured dialog tasks. The current trend towards generative models in various applications likely contributes to the increasing parameter sizes observed in recent models. Conversely, comprehension tasks, which typically involve more structured data, may not impose the same stringent demands for large parameter counts. This could explain why earlier BERT-like models, despite their relatively smaller parameter sizes, achieved strong performance in comprehension tasks, highlighting not only the advantage of model architecture but also the varying parameter requirements across different task types. BioQwen-1.5B, as expected, demonstrated its strength in generative tasks, outperforming the smaller model and, at times, approaching the performance levels of the larger BioQwen-1.8B, further reinforcing the notion that the unstructured nature of generative tasks benefits from larger parameter models.

In terms of efficiency, BioQwen-1.8B exhibited greater efficiency compared to BioQwen-0.5B, potentially due to its superior capacity for learning concise expressions in generative tasks. The inability of BioQwen-0.5B to directly answer English biomedical questions, a capability present in BioQwen-1.8B, further underscores the limitations associated with extremely small-parameter models. Notably, despite its challenges with comprehension tasks, BioQwen-1.5B retains the ability to answer English biomedical questions.

Small-parameter LLMs are expected to further enhance their comprehension and generation capabilities, showing significant potential in applications such as biomedicine and mobile devices [52,53]. Efficient quantization strategies and automated compilation workflows are likely to reduce computational and storage costs, enabling lightweight solutions for resource-constrained environments [54,55]. However, mobile deployment faces dual challenges: only flagship chipsets, such as Snapdragon 8 Gen 3 and 8 Gen 4, support smooth inference, while adapting models to platforms like iOS or Harmony OS requires substantial customized development due to architectural and memory management differences [56].

Despite these challenges, the potential of small-parameter LLMs in multi-task and multi-language scenarios is substantial. Applications like medical diagnosis support and intelligent customer service particularly benefit from their lightweight design and low latency. By further optimizing computational efficiency and domain-specific fine-tuning while maintaining accuracy [57], small-parameter LLMs can address current performance bottlenecks and become vital tools in real-world production environments.

However, this study has certain limitations. Our evaluation primarily focused on specific biomedical datasets. The consistent underperformance of BioQwen-1.5B in comprehension tasks, despite performing well in generative tasks, warrants further investigation into the reasons behind this discrepancy. The inability of BioQwen-0.5B to directly answer English biomedical questions also remains a limitation.

Future research should further extend the application of the model to a wider range of biomedical tasks and datasets to validate our findings.

Additionally, exploring more efficient training strategies and model architectures will help reduce resource consumption while maintaining performance. Further investigation into the impact of different pre-training and fine-tuning methods on model performance will provide deeper insights into optimizing model training. By leveraging rich high-quality biomedical corpora and designing effective fine-tuning strategies, we have significantly enhanced the performance of small-parameter LLMs in the biomedical field. Future work will focus on further enhancing the task capabilities, interpretability, and safety of the BioQwen models in biomedical applications.

5. Conclusion

In this study, we developed a series of small-parameter biomedical bilingual multi-task models, BioQwen, in three versions: 0.5B, 1.5B, and 1.8B. The results demonstrate that BioQwen significantly improves the efficiency and accuracy of biomedical text mining, surpassing recently released large-parameter biomedical models such as Taiyi, MedChatZH, and HuaTuoGPT2 across various tasks.

The construction of BioQwen incorporated an innovative data preprocessing pipeline, combined with a two-stage fine-tuning strategy enhanced by QLoRA and FlashAttention-2. These approaches substantially improved BioQwen's capabilities in handling biomedical bilingual texts (Chinese and English) and expanded its zero-shot learning ability to native English biomedical QA tasks. This advancement provides robust support for acquiring cross-lingual biomedical knowledge. Furthermore, the successful deployment of the quantized version of BioQwen-1.8B on mobile devices underscores its practical applicability in resource-constrained environments.

In summary, the small-parameter BioQwen model demonstrates state-of-the-art efficiency and accuracy in biomedical text mining, particularly in bilingual contexts. This study provides a feasible methodology for efficiently training small-parameter models and highlights their potential for real-world applications, as evidenced by their successful deployment on mobile platforms. We anticipate that this work will advance the practical application of biomedical text mining technology and contribute to the broader development of cross-lingual and resource-efficient models in the biomedical domain.

CRediT authorship contribution statement

Yang Yinqi: Data curation. **Bai Mingze:** Writing – review & editing. **Pu Dan:** Writing – review & editing. **Xie Jiazheng:** Writing – review & editing. **Liu Chuan:** Writing – review & editing. **Li Bo:** Writing – review & editing. **Liu Mingwei:** Writing – review & editing. **Shu Kunxian:** Writing – review & editing. **Li Yinghong:** Writing – review & editing, Project administration. **Yan Yudong:** Writing – original draft, Validation, Methodology, Data curation. **Tong Zhuohao:** Data curation. **Wang Yu:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62101087), the China Postdoctoral Science Foundation (Grant No. 2021MD703942), the Chongqing Postdoctoral Research Project Special Funding (Grant No. 2021XM2016), the Science and Technology Research Program of the Chongqing Municipal Education Commission (Grant No. KJQN202100642), and the Open Project of Sichuan Provincial Key Laboratory of Central Nervous System Drugs Co-built by Department and Municipality (Grant No. 210017–01SZ).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.asoc.2025.113084.

Data availability

I have made the relevant data and code open-source. The link can be found in the 'Data Availability' section of the manuscript.

References

[1] J. Liu, C. Wang, S. Liu, Utility of ChatGPT in clinical practice, J. Med. Internet Res. 25 (2023) e48568.
[2] H. Lee, The rise of ChatGPT: exploring its potential in medical education, Anat. Sci. Educ. 2023 (2023) ase.2270.
[3] A. Nazir, Z. Wang, A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges, Meta-Radiol. 1 (2023) 100022.
[4] J. Kaplan, S. Mccandlish, T. Henighan, et al., Scaling laws for neural language models, ArXiv 2020 (2020).
[5] M. Abdin, S.A. Jacobs, A.A. Awan, et al., Phi-3 technical report: a highly capable language model locally on your phone, ArXiv 2024 (2024), 2404.14219.
[6] A. Radford, J. Wu, R. Child, et al., Proc. Open. Blog (2019).
[7] T.B. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, ArXiv 2020 (2020), 2005.14165.
[8] XAI-ORG, Grok-1: Grok open release. Editor, Book Grok-1: Grok open release, Series Grok-1: Grok open release, 2024.
[9] Y. Bai, A. Jones, K. Ndousse, et al., Training a helpful and harmless assistant with reinforcement learning from human feedback, ArXiv 2022 (2022), 2204.05862.
[10] W.-L. Chiang, L. Zheng, Y. Sheng, et al., Chatbot Arena: an open platform for evaluating LLMs by human preference, ArXiv abs/2403.04132 (2024).
[11] M. Reid, N. Savinov, D. Teplyashin, et al., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, ArXiv abs/2403.05530 (2024).
[12] Anthropic, Claude, (Version 2.0), Anthropic, 2024.
[13] H. Touvron, T. Lavril, G. Izacard, et al., LLaMA: open and efficient foundation language models, ArXiv abs/2302.13971 (2023).
[14] J. Bai, S. Bai, Y. Chu, et al., Qwen technical report, ArXiv 2023 (2023), 2309.16609.
[15] A. Yang, B. Xiao, B. Wang, et al., Baichuan 2: open large-scale language models, ArXiv 2023 (2023), 2309.10305.
[16] A. Dubey, A. Jauhri, A. Pandey, et al., The Llama 3 herd of models, ArXiv abs/2407.21783 (2024).
[17] G.T.M. Riviere, S. Pathak, P.G. Sessa, et al., Gemma 2: Improving Open Language Models at a Practical Size, ArXiv abs/2408.00118 (2024).
[18] L.B.A.a.A.L.a.E.B.a.G.M.B.a.L.T.a.A.P.a.A.M.a.C.Z.a.L.v.W.a.T. Wolf, SmolLM2 - with great data, comes great performance, 2024.
[19] Q. Chen, H. Sun, H. Liu, et al., An extensive benchmark study on biomedical text generation and mining with ChatGPT, Bioinformatics 39 (2023) btad557.
[20] L. Luo, J. Ning, Y. Zhao, et al., Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks, J. Am. Med. Inform. Assoc. 2024 (2024) ocae037.
[21] M. Alkaoud, A bilingual benchmark for evaluating large language models, PeerJ Comput. Sci. 10 (2024) e1893.
[22] K. Singhal, T. Tu, J. Gottweis, et al., Towards expert-level medical question answering with large language models, ArXiv 2023 (2023), 2305.09617.
[23] H. Zhang, J. Chen, F. Jiang, et al., Huatuogpt, towards taming language model to be a doctor, ArXiv 2023 (2023), 2305.15075.
[24] Y. Tan, Z. Zhang, M. Li, et al., MedChatZH: a tuning LLM for traditional chinese medicine consultations, Comput. Biol. Med. 172 (2024) 108290.
[25] J. Chen, X. Wang, A. Gao, et al., Huatuogpt-II, one-stage training for medical adaption of LLMs, ArXiv 2023 (2023), 2311.09774.
[26] Z. Du, Y. Qian, X. Liu, et al., Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2021.
[27] Y. Li, Z. Li, K. Zhang, et al., ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge, Cureus 15 (2023).
[28] H. Xiong, S. Wang, Y. Zhu, et al., DoctorGLM: fine-tuning your Chinese doctor is not a herculean task, ArXiv abs/2304.01097 (2023).
[29] J. Zhang, R. Gan, J. Wang, et al., Fengshenbang 1.0: being the foundation of chinese cognitive intelligence, ArXiv 2022 (2022), 2209.02970.
[30] Y. Ji, Y. Deng, Y. Gong, et al., Exploring the impact of instruction data scaling on large language models: an empirical study on real-world use cases, ArXiv abs/2303.14742 (2023).
[31] Y. Bai, X. Du, Y. Liang, et al., COIG-CQIA: quality is all you need for Chinese instruction fine-tuning, ArXiv 2024 (2024), 2403.18058.
[32] S. Gururangan, A. Marasović, S. Swayamdipta, et al., Don't stop pretraining: adapt language models to domains and tasks, ArXiv 2020 (2020), 2004.10964.
[33] J. Lee, W. Yoon, S. Kim, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[34] R. Luo, L. Sun, Y. Xia, et al., BioGPT: generative pre-trained transformer for biomedical text generation and mining, Brief. Bioinforma. 23 (2022) bbac409.

[35] S. Huang, M. Noukhovitch, A. Hosseini, et al., The N+ implementation details of RLHF with PPO: a case study on TL;DR summarization, ArXiv 2024 (2024), 2403.17031.
[36] C. Fourrier, N. Habib, A. Lozovskaya, et al., Open LLM Leaderboard v2. Editor, Book Open LLM Leaderboard v2, Series Open LLM Leaderboard v2, Hugging Face, 2024.
[37] T. Dettmers, A. Pagnoni, A. Holtzman, et al., QLoRA: efficient finetuning of quantized LLMs, ArXiv 2023 (2023), 2305.14314.
[38] T. Dao, FlashAttention-2: faster attention with better parallelism and work partitioning, ArXiv 2023 (2023), 2307.08691.
[39] S. Huang, T. Cheng, J.K. Liu, et al., OpenCoder: the open cookbook for top-tier code large language models, ArXiv abs/2411.04905 (2024).
[40] N. Zhang, Z. Bi, X. Liang, et al., CBLUE: a Chinese biomedical language understanding evaluation benchmark, ArXiv 2021 (2021) 2106.08087.
[41] W. Zhu, X. Wang, H. Zheng, et al., PromptCBLUE: a Chinese prompt tuning benchmark for the medical domain, ArXiv 2023 (2023), 2310.14151.
[42] R.I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.
[43] J. Li, Y. Sun, R.J. Johnson, et al., BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database 2016 (2016) baw068.
[44] Z. Hongying, L. Wenxin, Z. Kunli, et al., Proceedings of the Chinese Lexical Semantics: 21st Workshop, CLSW 2020, Hong Kong, China, May 28–30, 2020, Revised Selected Papers 21, 2021, pp. 652-664.
[45] K. Papineni, S. Roukos, T. Ward, et al., Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311-318.
[46] A. Mutton, M. Dras, S. Wan, et al., Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 344-351.
[47] C.-Y. Lin, Proceedings of the text summarization branches out, 2004, pp. 74-81.
[48] J. He, M. Fu, M. Tu, Applying deep matching networks to Chinese medical question answering: a study and a dataset, BMC Med. Inform. Decis. Mak. 19 (2019) 91–100.
[49] S. Zhang, X. Zhang, H. Wang, et al., Multi-scale attentive interaction networks for Chinese medical question answer selection, IEEE Access 6 (2018) 74061–74071.
[50] T. Chen, T. Moreau, Z. Jiang, et al., Proceedings of the USENIX symposium on operating systems design and implementation, 2018.
[51] Q.A. Yang, B. Yang, B. Zhang, et al., Qwen2.5 technical report, ArXiv (2024).
[52] Z. Lu, X. Li, D. Cai, et al., Small language models: survey, measurements, and insights, arXiv preprint arXiv:2409.15790 (2024).
[53] C. Van Nguyen, X. Shen, R. Aponte, et al., A survey of small language models, arXiv preprint arXiv:2410.20011 (2024).
[54] K. Wang, Z. Liu, Y. Lin, et al., Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[55] M. Tan, Q.V. Le, EfficientNetV2: smaller models and faster training, arXiv preprint arXiv:2104.00298 (2021).
[56] A.G. Howard, M. Zhu, B. Chen, et al., MobileNets: efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).
[57] K.S. Phogat, S.A. Puranam, S. Dasaratha, et al., Fine-tuning smaller language models for question answering over financial documents, arXiv preprint arXiv: 2408.12337 (2024).