Quark

2025-8-19

QuarkMed Medical Foundation Model Technical Report

Ao Li[1], Bin Yan[1], Bingfeng Cai[1], Chenxi Li[1], Cunzhong Zhao[1], Fugen Yao[1], Gaoqiang Liu[1], Guanjun Jiang[1], Jian Xu[1], Liang Dong[1], Liansheng Sun[1], Rongshen Zhang[1], Xiaolei Gui[1], Xin Liu[1], Xin Shang[1], Yao Wu[1], Yu Cao[1], Zhenxin Ma[1] and Zhuang Jia[1]

[1]Quark Medical Team, Alibaba Group

arXiv:2508.11894v1 [cs.AI] 16 Aug 2025

Recent advancements in large language models have significantly accelerated their adoption in healthcare applications, including AI-powered medical consultations, diagnostic report assistance, and medical search tools. However, medical tasks often demand highly specialized knowledge, professional accuracy, and customization capabilities, necessitating a robust and reliable foundation model. QuarkMed addresses these needs by leveraging curated medical data processing, medical-content Retrieval-Augmented Generation (RAG), and a large-scale, verifiable reinforcement learning pipeline to develop a high-performance medical foundation model. The model achieved 70% accuracy on the Chinese Medical Licensing Examination, demonstrating strong generalization across diverse medical benchmarks. QuarkMed offers a powerful yet versatile personal medical AI solution, already serving over millions of users at https://ai.quark.cn.

1. Introduction

The advent of large language models (LLMs) has marked a pivotal moment in artificial intelligence, demonstrating remarkable capabilities in understanding and generating human-like text across a multitude of domains. This progress has catalyzed significant interest in their application to specialized fields, particularly medicine, where they hold the potential to revolutionize medical information retrieval, enhance early diagnostic accuracy, and support personalized health care requirements.

However, the medical domain presents unique and formidable challenges [47]. Unlike general-domain text, medical language is characterized by a highly specialized vocabulary, complex clinical concepts, and a nuanced syntax that is often ambiguous and context-dependent. As a result, general-purpose LLMs, which are typically fine-tuned for broad, non-medical corpora, often lack the deep, specialized knowledge required for high-stakes medical applications [1]. This knowledge gap can lead to unsatisfactory, and at times unsafe, performance when these models are directly applied to medical tasks.

Recognizing these limitations, the research community has shifted towards developing domain-specific foundation models for medicine. This endeavor began with the adaptation of Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers). Early pioneering work led to the creation of models such as BioBERT [24], which was pre-trained on large-scale biomedical literature, and ClinicalBERT [18], which was trained on unstructured clinical notes from electronic health records (EHRs). These models demonstrated that domain-specific pre-training significantly improves performance on various biomedical text mining tasks. Following this trend, models like BEHRT were developed to specifically model structured EHR data for predicting clinical events [27].

The success of these earlier models paved the way for the development of generative models tailored for medicine. BioGPT, for instance, was a generative pre-trained transformer that excelled at creating fluent biomedical text and improving performance on downstream tasks [29]. As model scaling became a key driver of performance, the field saw the emergence of significantly larger and more powerful medical LLMs. Models like GatorTron, with billions of parameters trained on massive clinical text datasets, demonstrated the benefits of scale in capturing the long-range dependencies and intricate relationships within clinical narratives [49].

More recently, the landscape has been defined by even larger and more sophisticated models that integrate extensive medical knowledge with robust instruction-following capabilities. Med-PaLM and its successor were among the first to approach expert-level performance on medical licensing examination-style questions, leveraging a combination of improved base models, medical domain fine-tuning, and advanced prompting strategies [1, 39]. Concurrently, the open-source community has produced a variety of powerful medical LLMs. Models like PMC-LLaMA [45], MEDITRON-70B [7], BioMedLM [33], and BioMistral [23] have been developed by pre-training on vast corpora of biomedical literature and clinical data, showing performance competitive with proprietary models. This proliferation of models has been accompanied by the creation of more comprehensive and challenging benchmarks, such as MedExQA [12] and MedS-Bench [46], which evaluate LLMs on more complex, long-form question answering and a wider array of clinical tasks.

Beyond supervised learning, Reinforcement Learning (RL) has emerged as a powerful paradigm for optimizing sequential decision-making, making it a promising approach for healthcare applications such as developing dynamic treatment regimes [19]. Concurrently, Reinforcement Learning from Human Feedback (RLHF) is being explored to incorporate the nuanced expertise of clinicians into the training loop. This allows models to learn from expert guidance, helping to create AI systems that are better aligned with human values and clinical best practices, ultimately enhancing the safety and reliability of the next generation of medical foundation models. However, applying RL in medicine is fraught with challenges, including the need for vast amounts of high-quality data, the difficulty in defining accurate reward functions, and ensuring the interpretability and safety of the models [35].

To address these limitations, recent advancements have focused on making RL more reliable and verifiable. One such advancement is Reinforcement Learning with Verifiable Rewards (RLVR), which trains models using objective feedback where the correctness of an output can be unambiguously determined, thereby mitigating the risk of "reward hacking" [16]. While RLVR has been effective in fields like mathematical reasoning , research is underway to adapt it to the complexities of medicine, where simple binary verification is often insufficient.

While these advancements are significant, there remains a critical need for medical foundation models that are not only knowledgeable and accurate but also highly reliable and customizable for real-world medical applications. This report introduces QuarkMed, a medical foundation model designed to meet these demands. By leveraging meticulously curated medical data, advanced Retrieval-Augmented Generation (RAG) for verifiable and up-to-date information, and a multi-stage training process including large-scale reinforcement learning, QuarkMed aims to provide a robust and versatile solution for a new generation of AI-powered healthcare tools.

We summarize the main contributions of this work:

- **Comprehensive Medical Data Pipeline:** A multi-layer curation and quality enhancement system (materials, structured knowledge, clinical records) with expert-guided coverage tracking and knowledge synthesis.
- **Ability- and Problem-Driven Instruction Tuning:** A multi-task IFT/SFT curriculum with automated ratio optimization and robustness-oriented adversarial augmentation.
- **Dual-Stage Reinforcement Learning:** A reasoning-focused verifiable reward phase followed by general alignment using multi-dimensional reward (honesty, helpfulness, consistency, compliance) and GRPO in medical domain.
- **Integrated Medical RAG:** Dense retrieval over authority-ranked corpora yielding large factuality and hallucination reductions with citation support.
- **State-of-the-Art Performance:** Strong results across public and internal medical exams (e.g., 70% Chinese Medical Licensing style accuracy) at a competitive 32B scale.

2. Data

On top of a general-purpose large language model, challenges remain in the model's parametric knowledge in specific medical domains, which can undermine the performance of downstream tasks. To bridge this gap, we employ a large-scale data processing pipeline to systematically prepare and ingest this knowledge into our base model. Three main types of medical-related data were used during different stages of model training: medical materials, medical knowledge, and medical records. This data contributes to the timely medical knowledge and detailed clinical knowledge for the model, complementing search-based augmented methods.

2.1. Medical Materials

To enhance our model's medical expertise, we have collaborated with an internal team of medical experts to build a large-scale, high-quality, and diverse dataset of medical materials. We also employ data synthesis to supplement knowledge points in critical areas.

Data Coverage and Scope Through various means such as web crawling and procurement, we have collected a wide range of data including textbooks, clinical guidelines, consensus statements, academic literature, drug inserts, medical encyclopedias, and clinical pathways. This effort has established a comprehensive medical materials library that provides approximately 1T tokens for model training. Based on a framework manually curated by medical experts, we implemented a fine-grained knowledge point coverage detection system. Drawing inspiration from Bloom's Taxonomy, we classify knowledge points into factual, conceptual, and procedural categories. Each category is evaluated using test sets built from Quark search query-cot and our internal medical knowledge graph. Through an iterative process of evaluation and supplementation, our final library achieves over 90% coverage for factual knowledge, 84% for conceptual knowledge, and 75% for qcot coverage. This progressive coverage, from foundational data to complex reasoning data, aligns with the different stages of model training.

Data Quality Enhancement A significant portion of the materials exists in image format. Initially, we used OCR and layout analysis models to extract text. To further improve extraction from images with complex layouts or backgrounds, such as those in popular science materials, we trained a fine-grained content structuring model based on qwen2.5 vl. This advanced approach improved the quality of the pre-training corpus by over 30% compared to the original OCR methods, achieving an average data usability rate of over 90%, with the rate approaching its upper limit for well-structured images like those in textbooks.

Authoritativeness and Verification To ensure the accuracy and reliability of the data, we established an authoritativeness labeling strategy based on the principles of evidence-based medicine. Based on factors such as the material's type, source, and impact factor, we classify data into five authority levels (A-E). This classification is used for filtering data in different stages, such as training and RAG. Within our library, high-authority data accounts for over 40% of clinical guidelines, 26% of literature, and 5% of books.

Knowledge Synthesis for Conceptual Gaps The content of original source materials may not adequately address certain high-level conceptual knowledge points. To address this, we employ data synthesis in key medical subdomains—such as diseases, symptoms, drugs, procedures, and tests—by systematically creating knowledge points for a nearly exhaustive set of entities across important relationships defined in our terminology sets. For instance, using data from drug regulatory agencies, we merge the inserts for the same generic drug from various manufacturers, combine this with pharmacological information from encyclopedias and textbooks, and synthesize a comprehensive insert for each generic drug name to be used for general knowledge enhancement.

2.2. Medical Knowledge

In the field of healthcare, it is crucial for language models to incorporate a certain level of medical background knowledge for the following reasons:

- **Improving Accuracy and Reducing Hallucinations:** A lack of professional expertise may lead to semantic confusion, resulting in inaccurate or misleading outputs. Incorporating domain-specific knowledge helps ensure that model predictions are grounded in established medical facts.
- **Enhancing Reasoning Capabilities:** Knowledge integration enables the model to perform more sophisticated reasoning tasks, such as inferring potential diseases from a set of symptoms or recommending appropriate diagnostic procedures.
- **Compensating for Gaps in Pre-training Corpora:** Certain critical information—such as data on rare diseases, newly developed drugs, or the latest clinical guidelines—may not be adequately represented in general-domain pre-training datasets. Supplementing with structured medical knowledge helps bridge these gaps and improves the model's relevance and applicability in real-world clinical settings.

Medical Data and Unstructured Data Processing The knowledge integrated into the training of large medical language models comes from multiple structured and unstructured sources, categorized based on content type and usage scenario. The classification and approximate scale of each category are summarized in Table 1. For unstructured data, it is used in stages such as continued pre-training, instruction fine-tuning, supervised fine-tuning, and reinforcement learning, based on method-specific data selection processes.

Table 1 | Classification and Scale of Medical Data Sources

Main Category	**Subcategories**	**Scale**
Web-based Resources	Q&A platforms, articles, encyclopedic entries	Tens of millions
Professional Materials	Clinical guidelines, publications, drug inserts, medical standards, medical exams	Millions
Knowledge Bases	Standard medical terminology sets, medical ontologies, dictionaries	Tens of millions
Medical Scenario Data	Online consultation dialogues, patient case records	Tens of millions
Other Supporting Data	Legal regulations, medical AI-related databases, clinical trial databases, doctor-patient communication data	Millions

Knowledge Transformation for Structured Data Since LLMs cannot directly use structured data, we transform it into natural language data using a knowledge transformation technique. The process of knowledge selection and construction follows a set of key criteria to ensure its effectiveness in supporting medical language understanding and reasoning. The selected knowledge must meet the following standards:

- **Importance:** The knowledge covers core medical concepts and relationships that are clinically significant.
- **Completeness:** It provides comprehensive coverage of relevant domains and avoids critical omissions.
- **Accuracy and Clarity:** Information is precise, well-defined, and free from ambiguity.
- **Generalization Ability:** The knowledge supports not only direct retrieval but also logical inference and reasoning over unseen or complex scenarios.

To align structured knowledge—such as Subject-Predicate-Object (SPO) triples from knowledge graphs—with the model's native capabilities, knowledge translation techniques are employed. These techniques convert structured SPO data into unstructured natural language sentences, ensuring compatibility with the model's processing pipeline and enabling more effective integration and learning. This process includes the following key steps:

- **Training the Translation Model:** A translation model is trained to map structured SPO triples into fluent, semantically equivalent natural language sentences. This involves constructing a parallel corpus of SPO triples and their corresponding textual descriptions and training a sequence-to-sequence model to learn the mapping.
- **Extracting Triples from Text and Performing Back-Translation:** To enrich the knowledge corpus and verify its consistency, the system performs triple extraction from unstructured medical texts and back-translation, where the extracted triples are re-translated into natural language to assess their coherence and correctness.
- **Quality Filtering:** A robust quality filtering mechanism is essential to maintain high data standards. This includes semantic consistency checks, grammatical and fluency evaluation, and domain relevance filtering.

Evaluation of Knowledge Integration To evaluate the effectiveness of this data, we use single-shot methods to probe the parametric knowledge in the model.

- **Knowledge Probes:** Knowledge probes are structured queries used to examine how well a model represents specific factual or conceptual knowledge. These probes help determine whether the model has not only memorized but also semantically understood the knowledge during training. Two types of test sets are constructed for comprehensive evaluation: a fact-based test set and a concept-based test set.
- **Query Optimization with Leading Text:** To improve the model's ability to retrieve relevant knowledge from its internal representations, we introduce "leading text"—a form of prompt engineering where additional contextual cues are prepended to the query. For example: "Based on the medical knowledge you have learned, what is the most likely diagnosis for a patient presenting with..."

The introduction of knowledge injection significantly improves the model's performance on both fact-based and concept-based knowledge probes. Specifically, accuracy increases from 39% to 60.57%, indicating a substantial enhancement in the model's ability to access and reason over the injected knowledge. The improvement is particularly notable in concept-based tasks, suggesting that the model has developed a better-structured understanding of medical domains. These results demonstrate that knowledge injection, when effectively integrated and evaluated through targeted probing methods, can significantly enhance the reasoning and knowledge utilization capabilities of large medical language models.

2.3. Medical Records

Real-world, high-quality medical records are invaluable for training medical foundation models, providing authentic clinical narratives, diagnostic reasoning, and treatment plans rarely captured in textbooks or guidelines. In this work, we curate a large-scale corpus from two practice-proximal channels: public online medical consultation dialogues and de-identified electronic health records (EHRs) released via public repositories. For confidentiality and compliance, we do not disclose exact dataset names or volumes; only aggregated, de-identified, privacy-filtered text is retained for modeling.

Online Medical Consultations We collect publicly available consultation dialogues that reflect symptom narratives, clinician questioning, preliminary differentials, and triage or follow-up suggestions. These short- and medium-form interactions complement formal records by capturing colloquial descriptions, lay terminology, and pragmatic decision cues encountered in routine care.

Public EHR Collections We draw on de-identified EHR datasets made publicly accessible through established releases, spanning both outpatient and inpatient encounters. Exact counts are intentionally abstracted (reported internally only) to reduce re-identification risk; qualitatively, coverage spans common ambulatory presentations through complex inpatient trajectories.

- **Outpatient EHRs:** Emphasize common presentations (chief complaint, history of present illness, assessment/plan, prescriptions) with broad breadth rather than disclosing absolute volume.
- **Inpatient EHRs:** Include admission notes, longitudinal progress notes, procedure and operative narratives, discharge summaries, laboratory panels, and imaging impressions across a wide diagnostic mix.

All materials—whether originally released in de-identified form or derived from public dialogues—undergo a conservative PHI-removal pipeline. We further normalize and segment unstructured text into coherent clinical documents. Quality is enforced through automatic discriminator models and physician-led spot audits. The corpus is used for continued pre-training to capture the structure and lexicon of clinical documentation, and for supervised fine-tuning (SFT) to strengthen the model's reasoning in complex scenarios.

3. Method

This section details the multi-stage training methodology for QuarkMed, which includes Instruction Fine-Tuning (IFT), Supervised Fine-Tuning (SFT), and two distinct stages of Reinforcement Learning (RL) designed to instill specialized reasoning and general alignment.

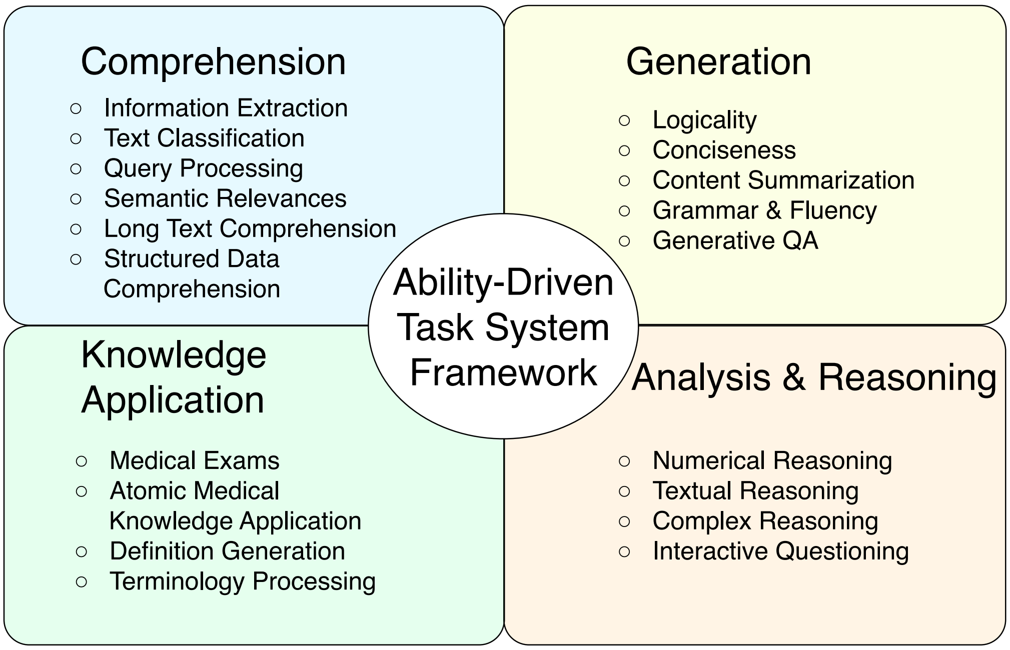

Figure 1 | Ability driven data augmentation loop

3.1. Instruction Fine-Tuning (IFT)

The initial phase of our training pipeline is Instruction Fine-Tuning (IFT), a critical step to align the general-purpose, pre-trained base model with the ability to follow instructions within the specialized medical domain. While pre-trained language models (PLMs) acquire extensive world knowledge through their auto-regressive training objective, they often fail to adhere to user-specific directives, a phenomenon known as the "alignment gap" [31, 1]. IFT seeks to bridge this gap by fine-tuning the model on a large and diverse dataset of instruction-response pairs. This process transforms the model from a simple text completion engine into a capable assistant that can understand and execute professional medical tasks. To this end, we have systematically constructed a task-oriented dataset comprising hundreds core IFT tasks with over 400,000 high-quality samples, following the principles of multitask-prompted training [36, 42].

Task Design Framework To construct a comprehensive instruction set, we adopted a dual-pronged strategy: a foundational, **Ability-Driven** framework and a responsive, **Problem-Driven** augmentation loop. This approach ensures broad capability coverage while systematically addressing identified model weaknesses. The Ability-Driven framework, illustrated in Figure 1, deconstructs the requirements of a medical AI assistant into a four-dimensional capability system.

- **Comprehension:** This dimension targets the model's fundamental understanding of prompts, including instructions, queries, and contextual information. We treat traditional Natural Language Understanding (NLU) tasks (e.g., information extraction, text classification, semantic similarity) as atomic abilities, which form the bedrock for performing more complex, domain-specific tasks.
- **Generation:** To ensure logical coherence, conciseness, and fluency, we designed tasks to refine specific aspects of text generation. For instance, a "Sentence Ordering" task enhances logical flow, while tasks for discriminating between hyponyms/hypernyms and synonyms reduce redundancy in enumerated lists.
- **Knowledge Application:** This dimension aims to activate the model's capacity to apply its parametric medical knowledge appropriately in different contexts. Tasks include adjusting terminology for diverse audiences (e.g., clinical vs. layperson) or applying knowledge of contraindications for specific populations (e.g., pediatric or geriatric patients) [38].
- **Analysis and Reasoning:** As the cornerstone of advanced medical AI, this dimension focuses on multi-step reasoning. We constructed a curriculum of reasoning tasks, beginning with atomic reasoning skills such as unit conversion and numerical comparison [8], and progressing to complex reasoning scenarios like multi-

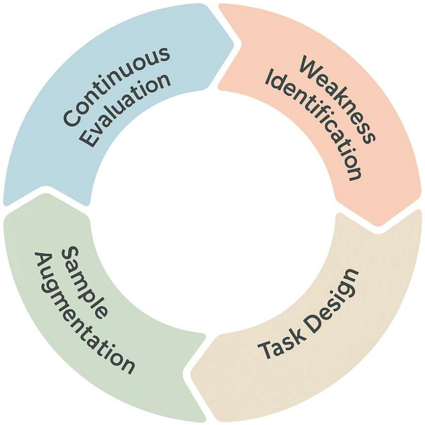

Figure 2 | Problem driven data augmentation loop

turn diagnostic dialogues and inference from clinical notes, inspired by chain-of-thought methodologies [44].

The Problem-Driven strategy establishes a continuous optimization cycle to address specific performance deficits identified during evaluation. As depicted in Figure 2, this iterative loop involves identifying model weaknesses, designing targeted tasks, and augmenting the training data to address these gaps. Key examples include:

- **Counterfactual Robustness:** To mitigate factual hallucinations, a "Factual Consistency Judgment" task was designed to train the model to identify and refuse to answer questions based on false premises.
- **Output Stability:** To improve robustness to linguistic variations, we generated "synonymous instruction - same output" pairs, ensuring the model produces consistent responses to semantically equivalent queries.
- **RAG Noise Resistance:** To enhance performance in Retrieval-Augmented Generation (RAG) scenarios [25], we constructed noisy samples containing both relevant and irrelevant retrieved passages. This trains the model to accurately identify, cite, and synthesize information from the most pertinent sources while ignoring distractors.

Guiding Principles for Task Construction The construction of our IFT dataset was governed by three core principles to ensure its effectiveness.

- **Task Atomicity:** Each task was designed to target a single, well-defined objective. This principle facilitates precise capability tracking and simplifies the attribution of performance changes, making the optimization process more controllable.
- **Instruction Generalization:** We developed unique prompt templates for each IFT task. This approach encourages the model to learn the underlying instruction-following behavior rather than memorizing surface-level patterns, thereby enhancing generalization to unseen tasks and isolating these foundational abilities from downstream, application-specific SFT.
- **Task Decomposition:** Complex, multi-step tasks that are difficult to learn end-to-end were systematically decomposed into more tractable sub-tasks. For example, in RAG scenarios, we created a distinct "Relevance Extraction" task to train the model to first identify useful information before generating a final answer.

Data Sourcing and Sample Construction We employed a multi-faceted strategy to generate high-quality training data tailored to the demands of each task.

- **High-Quality Base Samples:** For foundational NLU and generation tasks, we established a gold-standard reference by sampling outputs from multiple large models, followed by a cross-validation and voting process, and concluding with manual verification by domain experts.
- **Complex and Adversarial Samples:** For challenging tasks such as complex reasoning and counterfactual handling, we utilized a Self-Instruct approach [43], providing few-shot exemplars to guide a large model in generating a diverse set of novel prompts and responses.
- **Safety Alignment Samples:** To improve robustness against misuse, we trained a dedicated "Red-Teaming" model to generate adversarial prompts, enabling us to fine-tune the model for safer and more harmless responses, in line with established safety protocols [14, 2].

Training Strategy The IFT phase employed a sophisticated training strategy to maximize efficiency and model performance.

- **Curriculum Learning:** Recognizing that tasks exhibit varying levels of difficulty and that certain abilities are prerequisites for others (e.g., information extraction precedes summarization), we adopted a curriculum learning strategy [3]. Training progressed from simpler, atomic tasks to more complex, composite ones, which improved both convergence speed and final model performance.
- **Task Sample Ratio Optimization:** A key challenge is determining the optimal sampling ratio across the 112 tasks to achieve a balanced set of capabilities. We addressed this using a data-driven optimization process. First, we established automated evaluation suites for each core ability. We then used Bayesian Optimization, modeling the relationship between inter-group sampling ratios and a weighted overall performance score with a Gaussian Process Regression (GPR) model. An infill criterion was used to efficiently explore the high-dimensional search space and identify a near-optimal ratio distribution [13].

Through this systematic IFT process, the model's foundational abilities and instruction-following fidelity were significantly enhanced, providing a robust starting point for subsequent stages of fine-tuning and reinforcement learning.

3.2. Supervised Fine-Tuning (SFT)

The development of a safe, accurate, and helpful medical large language model hinges on the quality of its Supervised Fine-Tuning (SFT) data. To this end, we have designed a meticulous data processing pipeline that ensures our training samples are diverse, robust, and medically sound. The general process for the high-quality SFT samples generation is shown in Figure 3.

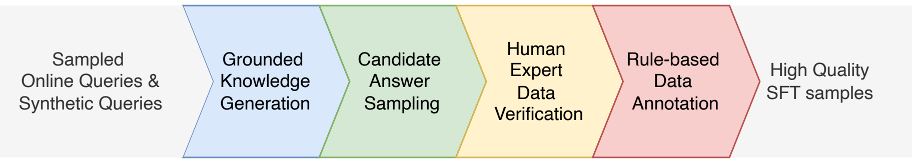

Figure 3 | Data processing pipeline for SFT samples

Query Selection Process Our strategy for Supervised Fine-Tuning (SFT) is built on a hybrid approach, utilizing both synthetic and real-world online data to create a medically proficient and adaptable AI. The primary goal is to develop a comprehensive set of skills, ranging from fundamental comprehension to advanced clinical reasoning. Synthetic data is systematically generated to train and test the model across a wide spectrum of controlled scenarios, establishing a strong foundation of core capabilities. This is complemented by data from online sources, which attunes the model to the nuance and diversity of real-world user queries, ensuring its practical applicability.

A significant focus of our training is on cultivating advanced comprehension and reasoning abilities. The model is trained to summarize lengthy and complex documents, synthesize key information scattered across multiple texts, and perform logical inference to connect disparate points. This training also involves handling information from diverse genres, such as clinical guidelines, Q&A forums, and patient dialogues. Furthermore, the model is taught to resolve conflicts by selecting the most accurate information from contradictory sources and to integrate knowledge from external tools, like BMI calculators, to provide more comprehensive answers. Table 2 shows the key capabilities for the synthetic data categories.

Ensuring safety and accuracy is paramount, and the model undergoes rigorous training to enhance its robustness against erroneous or misleading information. It is specifically trained to identify and resist various forms of disturbance, such as ignoring irrelevant reference materials, differentiating between semantically similar but incorrect medical concepts, and flagging factually incorrect information. This error resistance extends to the user's query itself, enabling the model to recognize and handle questions that contain flawed premises, factual inconsistencies, or irrational assumptions, thereby preventing the generation of unsafe or nonsensical responses.

Table 2 | Key Capabilities for SFT Synthetic Data Strategy

Capability Category	Specific Capability	Purpose
Summarization & Induction	Dispersed Information Synthesis	To combine scattered information into a coherent answer.
Disturbance Resistance	Contradictory Information	To identify and use correct information from conflicting sources.
Error Resistance	Factual Inconsistency	To recognize and handle incorrect or illogical user queries.
Fundamental Capabilities	Timeliness	To provide the most current and updated information.
	Authoritativeness	To learn to prioritize and cite authoritative sources.

Finally, this training is grounded by a focus on foundational skills and real-world performance. The model is explicitly taught to ensure its responses are logically structured, use the most current and timely information, and cite authoritative sources to build user trust. Capabilities such as correcting misspelled drug names and maintaining coherent, non-repetitive conversational flow are also instilled. By incorporating online data from clinical and general health contexts, we ensure these foundational skills are effectively applied to address genuine user needs, from complex clinical questions to practical advice on healthy living, while upholding strict safety standards against adversarial inputs.

Data Curation Process Our data curation process consists of four main stages: medical knowledge grounded generation, candidate answer sampling, human-expert data verification, and rule-based data annotation.

- **Medical Knowledge Grounded Generation**: To generate responses grounded in medical knowledge, we begin by sampling anonymous, real-world online queries. We leverage the powerful retrieval capabilities of Quark Medical Search to gather a comprehensive set of reference materials. This includes professional medical literature and medical question-and-answer forums, supplemented with timely, web-wide content such as informational notes. The objective is to create a SFT dataset that reflects real-world complexity, possesses a sufficient level of difficulty, and covers a wide spectrum of scenarios across medical knowledge, clinical practice, and medical application. All original data sources, such as proprietary medical texts and patient records, undergo rigorous privacy protection measures.
- **Candidate Answer Sampling**: For each prompt, we generate multiple candidate answers tailored to the specific requirements of each data sample. Utilizing a proprietary medical quality model alongside the capabilities of in-house reward models, we employ a "Best-of-N" strategy to select the most optimal response to serve as a candidate for the next stage.

- **Human-Expert Data Verification**: We assemble a team of medical experts to meticulously review and refine the candidate answers. This verification process ensures that every response adheres to our stringent standards of safety, accuracy, and usefulness. To enhance the quality and efficiency of this stage, annotators follow a structured format that includes the reference materials, the initial question, the model's best-selected answer, responses from other major models, and a summary of key points potentially required in the final answer.
- **Rule-Based Data Annotation**: In the final stage, we apply rule-based data annotations. This involves real-time services that provide automated validation of formatting and correctness, further elevating the overall quality and consistency of the annotated data.

Ultimately, this comprehensive process, which combines systematic query design with a multi-stage curation pipeline leveraging both automated systems and human medical expertise, ensures the creation of high-quality SFT data. This meticulous approach is fundamental to developing a trustworthy and capable medical AI assistant that is both safe and genuinely helpful to users.

3.3. Stage 1 RL: Large-Scale Medical Reinforcement Learning

Fields like medicine—especially core tasks such as disease diagnosis, rational drug use, and test ordering—are inherently knowledge-intensive and rely heavily on sophisticated reasoning. To significantly elevate the QuarkMed model's reasoning capabilities in these complex scenarios, we implemented a specialized Reinforcement Learning (RL) phase focused exclusively on reasoning-based tasks. By using a multi-task learning approach across medical board exams, disease diagnosis, appropriate medication prescribing, and lab/imaging test selection, our goal is to systematically enhance the model's overall medical reasoning competence.

Model Initialization with SFT To accelerate RL convergence and conserve computational resources, we begin with a "cold-start" strategy using Supervised Fine-Tuning (SFT). The objective is to ensure the model acquires a baseline reasoning capability and can adhere to predefined formats (e.g., JSON) before entering the more complex RL stage. We fine-tune the base model directly on a curated set of over 700 high-quality annotated examples from our target reasoning tasks. To preserve model diversity and prevent its entropy from dropping too low, we limit SFT to just two epochs. This encourages the model to "learn to reason" rather than "memorize answers," building a strong foundation for the exploration required in RL.

High-Quality Data Pipeline for RL Training The success of reinforcement learning hinges on the quality of its training data. Our data pipeline was designed around three core principles: diversity, difficulty, and accuracy. For **diversity**, we start with heterogeneous data sources, including electronic health records and medical exam questions, and use a label-based stratified sampling strategy to ensure the training data is balanced. For **difficulty**, we built a dynamic, model-aware filter to continuously feed the model challenging examples by screening for high-accuracy samples, analyzing reasoning complexity, and synthesizing more complex problems. This filtering pipeline was customized for each model architecture (Llama vs. Qwen) and size (8B vs. 32B). For **accuracy**, we use a "model-assisted, expert-verified" workflow where a high-performing model generates initial responses, discrepancies are flagged, and our team of medical experts reviews and corrects them to guarantee label accuracy.

Reward Model Design A precise reward signal is critical for guiding the RL training process. For reasoning tasks with a clear ground truth, we designed a hybrid reward model, encapsulated in a "Verifier," that prioritizes rules but is augmented by a model-based component. This approach is founded on a rule-first principle, where objective and stable reward signals from established medical rules are given precedence to prevent reward hacking. We also standardized output formats (e.g., ICD codes for diagnosis) to enable automated scoring. However, since simple rule-based matching can be brittle, we introduced a model-based reward to handle synonyms, hierarchies, and incomplete labels. To mitigate potential bias, this Verifier was iteratively optimized. Experiments confirmed that this hybrid "rule + model" strategy improved disease diagnosis performance by 3 percentage points over a purely rule-based method. Finally, a dedicated format-adherence reward is included to ensure outputs strictly follow the required structure. The training data and Verifier setup are summarized in the table below.

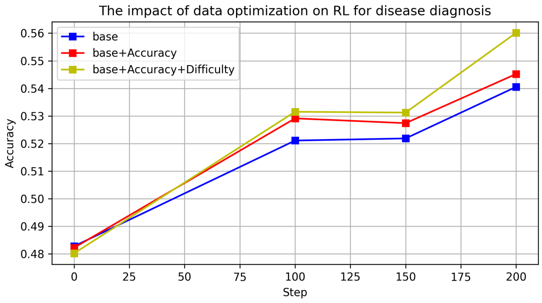

Figure 4 | This figure illustrates the impact of enhancing data accuracy and difficulty on RL performance. Both factors were found to positively influence the model's effectiveness in disease diagnosis tasks.

Table 3 | Task-Specific Verifier Composition and Key Metrics

Task Type	**Data Volume**	**Verifier Composition**	**Key Metrics**
Disease Diagnosis	16,000	Rules (ICD Matching) + Model	Accuracy, Recall
Rational Drug Use	10,000	Rules (JSON Extraction) + Model	Drug Entity Accuracy
Test Ordering	16,000	Rules (Keyword Matching)	Accuracy of Recommendations
Medical Exam Questions	27,000	Rules (Exact Answer Match)	Answer Accuracy

RL Implementation and Optimization We used the VeRL [37] framework for RL training, selecting the Group Relative Policy Optimization (GRPO) algorithm. GRPO normalizes the advantage function within groups of samples, which naturally supports multi-task training and improves stability. For efficiency and stability, we implemented several optimizations. First, dynamic resampling at the start of each epoch removes simple samples the model has already mastered, boosting training efficiency by about 20%. Second, to improve the stability of the policy model's exploration and the accuracy of the Verifier's scores, we increased the number of rollouts to 32 for the policy model and 8 for the Verifier. The final stage 1 model's performance on our test sets is shown below, demonstrating substantial improvements over the baseline.

Table 4 | Stage 1 Model Performance Comparison

Model	**Chinese National Medical Licensing Examination**				**Disease Diagnosis**	
	Junior	**Intermediate**	**Assoc. Senior**	**Senior**	**Top-1 Acc.**	**List Score**
DeepSeek-R1	0.814	0.826	0.723	0.387	0.75	1.46
Quark Stage1	0.822	0.772	0.683	0.524	0.86	3.32

3.4. Stage 2 RL: General Reinforcement Learning Integration

The primary objective of the general RL stage is to employ Reinforcement Learning (RL) to align the model's behavior with human preferences and values. This process involves developing a Reward Model (RM) to quantitatively assess the quality of model outputs and implementing an RL algorithm to optimize the policy based on these reward signals. To ensure the model generates high-quality medical responses, our RM holistically evaluates model outputs across three core dimensions: Honesty, Helpfulness, and Content Compliance.

Honesty Reward The Honesty Reward is designed to ensure the medical accuracy of the model's responses. To address the challenge of costly manual annotation for factual errors, we designed an iterative optimization loop involving a generative reward model model and the reward model, as show in Figure 6. First, we trained a generative reward model on manually calibrated SFT samples with Chain-of-Thought (CoT) reasoning and score. This model was then used to score multiple candidate responses and generate high-quality preference pairs.

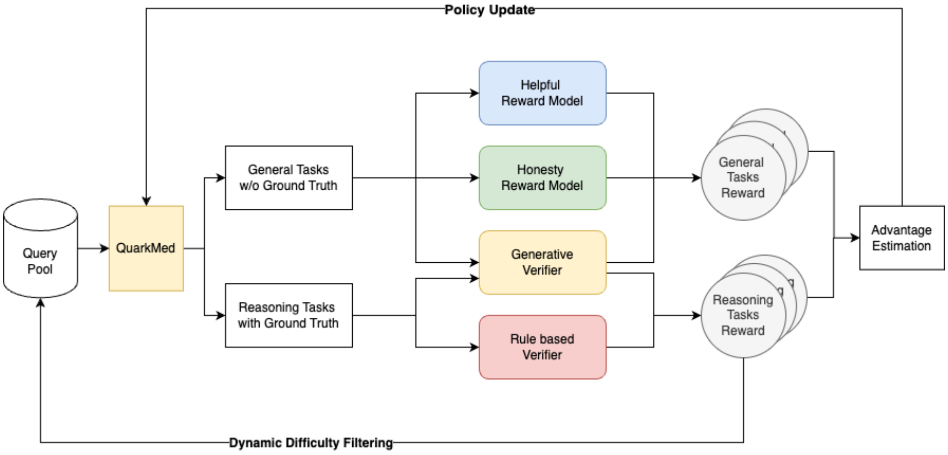

Figure 5 | Overall training method for the general RL stage. Three types of reward signals are used for general and reasoning tasks during the training.

Combining these filtered pairs with manually annotated samples, we trained the Reward Model (RM) using the Bradley-Terry (BT) model [4]. Finally, in a closed-loop iteration, the new RM evaluated more candidates, and samples with erroneous or ambiguous scores were fed back to human annotators for re-labeling, continuously enhancing both models.

Helpfulness Reward The evaluation criteria for helpfulness align closely with human intuition, making annotation relatively straightforward. We sampled tens of thousands of user prompts from online logs and used multiple models to generate diverse candidate responses. Annotators then provided preference labels on these varied outputs, which enhanced the generalization ability of the reward model. To counter "reward hacking" [15], where the policy learns to game the RM (e.g., by inflating response length), we established a continuous feedback loop, re-labeling new samples generated during RL training to iteratively update the RM and enhance its robustness.

Consistency Reward To address the issue of inconsistency between the reasoning process and the final summary in the model's outputs, we developed a dedicated consistency reward model. In order to improve model accuracy, we have constructed a multi-stage data iteration pipeline to enhance data quality. This pipeline incorporates key stages such as automated data collection via large model auto-labeling, model-based label calibration, and iterative active learning. Samples refined through this multi-stage process achieve a consistency reward score of over 80.

General Verifier for Content Compliance To meet strict style and formatting requirements for specific scenarios (e.g., health notes), we trained a General Verifier, which is an instruction-following model. By providing it with explicit evaluation principles, it can score responses based on adherence to these rules. This approach is highly flexible, as it can be adapted to new standards by modifying its guiding principles. It also effectively mitigates hacking, as new validation rules can be quickly added to address stylistic issues that are difficult for traditional RMs to cover.

Training Data Construction The training data for the RL stage comprises approximately 80,000 prompts, divided into reasoning-intensive and general-purpose categories. The reasoning-intensive data (approx. 60,000 prompts) was isolated from data used in earlier stages to ensure continued gains. The general-purpose data (approx. 20,000 prompts), sourced from online logs, was selected via diversity sampling. We used the SFT-stage model to perform multiple rollouts and prioritized prompts that yielded a high diversity in reward scores, indicating varied quality in model responses. Ultimately, we maintained a reasoning-to-general-purpose data ratio of approximately 3:1.

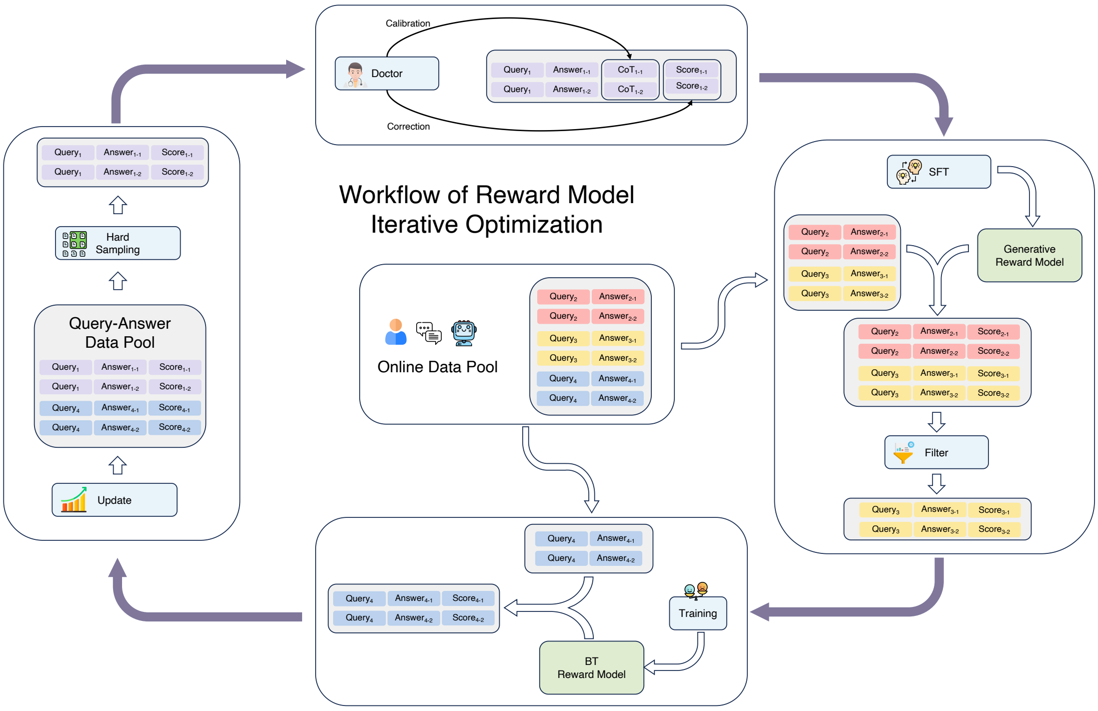
Workflow of Reward Model Iterative Optimization
Calibration
Doctor
Correction
Query$_1$ Answer$_{1-1}$ CoT$_{1-1}$ Score$_{1-1}$
Query$_1$ Answer$_{1-2}$ CoT$_{1-2}$ Score$_{1-2}$
SFT
Generative Reward Model
Query$_2$ Answer$_{2-1}$ Score$_{2-1}$
Query$_2$ Answer$_{2-2}$ Score$_{2-2}$
Query$_3$ Answer$_{3-1}$ Score$_{3-1}$
Query$_3$ Answer$_{3-2}$ Score$_{3-2}$
Filter
Online Data Pool
Query$_4$ Answer$_{4-1}$ Score$_{4-1}$
Query$_4$ Answer$_{4-2}$ Score$_{4-2}$
Training
BT Reward Model
Update
Query-Answer Data Pool
Hard Sampling

Figure 6 | Workflow of reward model iterative optimization. An iterative optimization loop that begins with manually labeled data to train a generative chain-of-thought (CoT) reward model. The CoT-reward model is used for pairwise selection to obtain high-quality pairs, which are then used to train a BT-reward model. The BT-reward model is further employed in an active learning setting to continuously improve both the training samples and the performance of the reward models.

Algorithm Comparison and Selection We conducted a comparative analysis of DPO (Direct Preference Optimization) [34] and GRPO (Group Relative Policy Optimization) to evaluate their effectiveness. For DPO, we formed preference pairs from the highest- and lowest-rewarded responses for each prompt. For GRPO, we sampled 8 candidate responses per prompt and applied the appropriate reward models or the General Verifier for scoring, with a KL divergence penalty coefficient of 0.01. Experimental results, shown in Table 5, demonstrated that GRPO significantly outperformed DPO across most dimensions. The GRPO model achieved the best performance in the overall score and on key dimensions such as Honesty and Harmlessness. Consequently, we selected GRPO as the core RL algorithm for this project.

Table 5 | Performance Comparison of RL Algorithms

Model / Algorithm	**Overall (5-pt)**	**Honesty (3-pt)**	**Harmless (3-pt)**	**Logicality (3-pt)**	**Formatting (3-pt)**	**Relevance (3-pt)**	**Comprehensive (3-pt)**
DeepSeek-R1 (Baseline)	3.60	2.24	2.76	2.80	2.92	2.94	2.60
DPO	3.49	2.16	2.72	2.66	2.88	2.92	2.72
GRPO	3.84	2.40	2.88	2.82	2.94	2.94	2.56

4. Evaluation

We conducted a comprehensive evaluation of the QuarkMed model using a wide range of benchmarks, including prominent open-source medical evaluation suites and several private datasets. We categorized the medical tasks into three main types: Medical Question Answering (QA), Medical Reasoning, and Foundational Medical

Capabilities. Tables 6 and 7 presents a summary of the results across these tasks and datasets, with detailed descriptions of each provided in the subsequent sections.

4.1. Evaluation Methodology

Due to data privacy and licensing constraints, all competitor models selected for comparison were either open-source models or accessible via publicly available APIs. During the evaluation, we performed a single inference pass for each test sample. Across all benchmark tests, a temperature parameter of 0.6 was consistently used for inference with every model. For certain datasets, such as MMLU [17], where test set answers are not provided, we conducted our evaluation on the validation set. For datasets with an excessively large number of test samples, such as RareBench [6], we performed uniform sampling, using a maximum of 1,000 samples per test set. We re-executed the entire evaluation pipeline for all selected models and APIs using a standardized prompt to ensure consistent results. For multiple-choice questions, models were prompted to provide the final answer in JSON format. For open-ended question formats, we employed DeepSeek-V3-0324 [10] for standardized post-processing and answer scoring.

4.2. Datasets

Medical Question Answering This category includes datasets such as MedQA [20], MedMCQA [32], PubMedQA [21], CMExam [28], and AfriMed-QA [30].

- **MedQA**: is an open-ended multiple-choice question dataset for the medical domain. In this paper, we only used the USMLE section data.
- **MedMCQA**: contains high-quality multiple-choice questions from the AIIMS and NEET PG entrance examinations, covering 2,400 medical topics and 21 medical subjects.
- **PubMedQA**: is a biomedical question-answering dataset constructed from PubMed abstracts. Given a medical research question, the model must answer "yes," "no," or "maybe" based on the provided abstract.
- **CMExam**: is derived from the Chinese National Medical Licensing Examination and contains over 60,000 multiple-choice questions for standardized medical assessment.
- **AfriMed-QA**: is a large-scale, open-source dataset featuring clinically diverse questions and answers from a Pan-African context. It is designed for the rigorous evaluation of large language models on accuracy, factualness, hallucination, demographic bias, potential harms, comprehension, and memory. We utilized only the multiple-choice portion of this dataset.

Medical Reasoning This category comprises datasets focused on disease reasoning and diagnosis, including DiagnosisArena [50], RareBench, MedXpertQA[51], and others.

- **MedXpertQA** is a highly challenging and comprehensive benchmark designed to evaluate expert-level medical knowledge and advanced reasoning capabilities.
- **DiagnosisArena** contains 1,000 paired, multi-turn patient case dialogues and their corresponding diagnoses, designed to evaluate the diagnostic reasoning capabilities of large language models in a clinical setting.
- **RareBench** is used to systematically evaluate the capabilities of large language models across four key dimensions within the rare disease domain. We used only the differential diagnosis section.
- **MedBullets** [5] is a medical question bank of simulated clinical problems, including 308 clinical multiple-choice questions in the style of USMLE Step 2 and Step 3.
- **CMB-clin** [41] based on complex real-world clinical cases, assesses a model's ability to apply knowledge in authentic diagnostic and treatment scenarios, evaluating whether it can leverage this knowledge to solve complex clinical problems.
- **CPQExam** is a private dataset built from the Chinese Health Professional Qualification Examination, an exam that physicians in China are required to pass for professional licensure and career advancement. The exam contains a large number of questions focusing on case analysis and practical application.
- **MediQ** [26] simulates interactive dialogues between patients and specialists. We used only the diagnosis-related multiple-choice questions, which require deriving a final diagnosis from the provided information.
- **RedisQA**[40] is a dataset built around rare disease diagnosis, covering 205 rare diseases to evaluate the performance of large language models in this area.

Foundational Medical Capabilities This category consists of the medical-related subsets of MMLU and the MedCalc [22] benchmark.

- **MMLU** is a multi-task benchmark featuring multiple-choice questions from various fields of knowledge. We selected only its medical-related subjects: virology, professional medicine, medical genetics, college medicine, college biology, clinical knowledge, and anatomy.
- **MedCalc** is a benchmark designed to assess a model's proficiency as a clinical calculator. Each sample includes a question that requires the model to calculate a clinical value based on a provided patient note.

Table 6 | Performance on Open Benchmarks.

	Large Models						**Medium-sized Models**			
Dataset	**Gemini-2.5-pro-0617**	**gpt-4o-2024-08-06**	**DeepSeek-R1-0528**	**Kimi-k2**	**Qwen3-235B-A22B-Instruct-2507**	**Qwen3-235B-A22B**	**o3-mini**	**Qwen3-32B**	**Baichuan-M1-14B**	**QuarkMed (Ours)**
MedQA(US)	**92.61%**	89.23%	90.02%	88.17%	84.02%	87.43%	74.46%	**87.19%**	69.83%	86.02%
MedMCQA	**82.73%**	76.90%	79.87%	78.01%	72.89%	76.79%	60.57%	71.31%	65.90%	**75.50%**
PubMedQA	76.40%	71.80%	73.20%	**77.25%**	75.20%	74.80%	73.60%	73.40%	70.80%	**79.00%**
CMExam	86.37%	78.88%	87.50%	88.79%	**90.10%**	86.90%	70.74%	85.80%	76.70%	**88.60%**
AfriMed-QA	85.57%	82.50%	**85.70%**	76.57%	76.64%	81.80%	**80.60%**	75.90%	67.20%	74.40%
MedXpertQA	**46.42%**	25.90%	39.30%	30.65%	32.63%	33.37%	**35.43%**	26.14%	23.00%	28.68%
DiagonissArena	**65.91%**	51.90%	60.65%	57.06%	54.38%	50.00%	56.00%	52.17%	51.50%	**61.90%**
RareBench	55.86%	55.97%	57.56%	50.25%	**57.98%**	49.24%	55.06%	50.51%	**56.46%**	52.90%
MedBullets	**82.24%**	76.30%	82.06%	75.53%	80.56%	78.66%	**83.66%**	74.14%	60.55%	77.27%
CMB-clin	3.52	3.17	3.56	**3.70**	3.57	3.50	**3.60**	3.50	3.07	3.50
MediQ	**95.22%**	86.70%	92.66%	84.66%	82.81%	92.55%	**90.78%**	87.17%	72.40%	85.06%
RedisQA	**88.54%**	82.30%	88.47%	85.71%	85.99%	85.81%	**87.35%**	82.10%	77.10%	83.20%
MMLU(Med)	**90.18%**	88.20%	89.23%	88.64%	88.73%	88.60%	87.01%	87.48%	81.43%	**88.37%**
MedCalc	**32.99%**	31.31%	29.96%	32.04%	28.38%	25.30%	35.78%	25.53%	**38.41%**	30.61%
Average*	**76.36%**	69.80%	74.66%	71.85%	71.40%	71.34%	70.07%	69.02%	63.43%	**71.36%**

* The 4-point scores from the CMB-clin dataset were normalized to a 0-1 range to calculate the overall mean.

Table 7 | Performance on Exam Benchmarks (CPQExam).

Classification Method	**Category**	**Gemini-2.5-pro-0617**	**o3-mini**	**gpt-4o-2024-08-06**	**DeepSeek-R1-0528**	**Kimi-k2**	**Qwen3-235B-A22B-Instruct-2507**	**Qwen3-235B-A22B**	**Qwen3-32B**	**Baichuan-M1-14B**	**QuarkMed (Ours)**
Subject	Primary Level	80.50%	70.50%	71.67%	81.42%	83.16%	83.03%	78.76%	78.76%	75.67%	81.50% (**83.3**%*)
	Intermediate Professional	79.87%	66.99%	64.15%	82.58%	74.71%	76.79%	72.42%	70.53%	68.04%	75.08% (**85.4**%*)
	Associate Senior Professional	65.29%	54.25%	43.92%	72.33%	55.42%	59.11%	56.16%	56.15%	50.33%	66.67% (**75.3**%*)
	Senior Professional	33.60%	35.50%	12.40%	38.70%	21.48%	27.45%	30.34%	35.18%	30.30%	51.70% (**67.7**%*)
Question type	Multiple-Choice	87.45%	74.74%	75.58%	88.14%	87.30%	87.80%	82.08%	79.74%	78.32%	82.58% (**91.80**%*)
	Multiple-Response	27.19%	37.15%	3.63%	32.89%	10.60%	17.21%	26.17%	33.90%	29.34%	55.72% (**76.40**%*)
	Shared Stem	80.06%	69.58%	63.86%	85.31%	76.06%	76.38%	71.34%	70.29%	59.64%	74.10% (**81.30**%*)
	Case Analysis	43.05%	36.82%	23.74%	48.75%	32.78%	39.47%	37.69%	38.64%	32.21%	49.85% (**58.50**%*)

* Knowledge augmentation was employed during the prediction phase.

4.3. Results

As shown in Table 6, the QuarkMed model demonstrates superior overall performance on medical benchmarks compared to Qwen3-32B [48], establishing it as one of the most powerful models in its size class. Notably, the QuarkMed model demonstrates exceptional performance on CPQExam, significantly surpassing other powerful models such as DeepSeek-R1-0528 [11], o3-Mini, and Gemini-2.5-pro-0617 [9] as shown in Table 7.

This result underscores the effectiveness of our Reinforcement Learning (RL) training tailored for medical scenarios and highlights the importance of domain adaptation. Compared to larger open-source models like Qwen-235B-A22B and Kimi-k2, QuarkMed also achieves superior performance on several reasoning datasets, such as MedXpertQA and DiagnosisArena, lagging only behind some prominent closed-source models such as Gemini-2.5-pro-0617. This indicates that our multi-stage training approach, grounded in medical domain knowledge, effectively enhances the model's performance on medical reasoning tasks.

5. Discussion

This section summarizes practical lessons, current limitations, and forward-looking directions for QuarkMed.

5.1. Enhancing Performance with Retrieval-Augmented Generation (RAG)

Although substantial effort was devoted to enriching the model's parametric (internal) medical knowledge, the strongest and most reliable performance in day-to-day medical assistance and exam-style question answering still hinges on RAG. For high-stakes factual, guideline-timed, or emerging-topic queries, parametric recall alone plateaus: subtle distinctions (e.g., edition-specific dosing updates, recency-dependent public health advisories, or differential refinements) benefit disproportionately from grounding in curated, authority-ranked external sources. In production usage we observe:

- Marked gains in factual precision and reduction of subtle hallucinations (e.g., obsolete regimen recommendations) when RAG is enabled, especially for infrequent entities and recently updated contraindications.
- Improved calibration: the model is more likely to qualify uncertainty or surface alternative hypotheses when contrasting retrieved passages diverge.
- Better exam robustness: for multi-step clinical vignettes, retrieved snippets often disambiguate near-miss distractors (pathophysiology nuance, epidemiologic prevalence shifts) that parametric memory alone confuses.

Thus, RAG acts not as an auxiliary enhancement but as a primary reliability layer, and investment priority (index freshness, authority scoring, redundancy pruning, and noise-resilient prompt packaging) remains critical to sustained performance.

5.2. Implications and Limitations of Reinforcement Learning

Reinforcement learning (RL) substantially improved structured reasoning (diagnosis selection, test ordering, medication rationality) when clear, automatable verifiers or semi-structured labels existed. Advantages observed:

- Reward shaping with hybrid rule+model verifiers amplified format fidelity and reduced reward gaming versus purely model-based preference signals.
- Group-wise normalization (GRPO) stabilized multi-task optimization under heterogeneous reward scales.
- Curriculum-style difficulty resurfacing prevented early convergence on shallow heuristics.

However, limitations persist:

- Verifiability Bias: Performance gains concentrate in domains with discrete, checkable endpoints (ICD codes, structured options). Nuanced counseling, longitudinal management planning, lifestyle tailoring, and patient education—where correctness is gradient, contextual, or preference-dependent—remain under-optimized.
- Reward Coverage Gaps: Current verifiers incompletely model temporal reasoning (trajectory forecasting, de-escalation strategies), causal justification coherence, and uncertainty articulation.
- Overfitting Risk: Tight coupling to deterministic format/verifier schemas risks brittle behavior when schema shifts (new coding versions, guideline reframing).
- Sparse / Delayed Feedback: Multi-turn dialogue quality, safety under adversarial probing, and cumulative patient-centric utility lack dense, reliable automatic signals.

- Alignment Trade-offs: Maximizing verifiable reasoning occasionally reduces stylistic empathy or brevity unless explicitly multi-objective tuned.

Future RL extensions need: (i) semi-verifiable composite rewards (fusing probabilistic factuality estimators with discourse/causal coherence scoring), (ii) active uncertainty elicitation (rewarding calibrated deferral or source citation), (iii) hierarchical RL separating strategic clinical framing from tactical entity selection, (iv) integration of simulation or synthetic patient state transitions for temporal credit assignment, and (v) continual verifier refresh pipelines aligned with evolving guidelines.

5.3. Future Directions

Despite its strong performance, the development of QuarkMed highlighted several challenges that point toward future research directions. A key challenge remains the dynamic and ever-evolving nature of medical knowledge. While our RAG system helps, ensuring real-time updates and resolving conflicts between different sources is an ongoing effort. Another limitation is the model's current focus on text-based data. Future work will concentrate on developing multi-modal capabilities, enabling QuarkMed to interpret medical images such as X-rays and pathology slides, which are crucial for many diagnostic workflows.

Furthermore, we aim to improve real-time personalization. Tailoring medical information to an individual's specific health context, while strictly preserving privacy, could significantly enhance the utility of our AI assistant. Finally, we will continue to refine our verification and citation mechanisms. Improving the granularity of citations and developing more robust methods for the model to self-correct and express uncertainty will be critical for building a safer and more reliable medical foundation model.

6. Conclusion

This report introduces QuarkMed, a 32-billion parameter foundation model specifically designed for the medical domain. We have detailed our comprehensive, multi-stage approach, which begins with curating a massive and diverse corpus of high-quality medical data. The training methodology combines Supervised Fine-Tuning with two distinct stages of Reinforcement Learning: one focused on verifiable, reasoning-intensive tasks and another on general alignment with human preferences for safety and helpfulness.

The effectiveness of this approach is demonstrated by QuarkMed's state-of-the-art performance on both public and internal benchmarks, including a 70% accuracy on the Chinese Medical Licensing Examination. By integrating an advanced Retrieval-Augmented Generation system, QuarkMed ensures its responses are grounded in timely and authoritative medical knowledge. As a powerful and versatile AI solution already serving millions of users, QuarkMed represents a significant step forward in developing reliable and effective AI tools for healthcare. We believe this work contributes to the broader goal of leveraging artificial intelligence to improve access to medical information and support better health outcomes globally.

References

[1] S. Azizi, V. Natarajan, C. Semturs, and et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[2] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

[4] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[5] H. Chen, Z. Fang, Y. Singla, and M. Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.182. URL https://aclanthology.org/2025.naacl-long.182/.

[6] X. Chen, X. Mao, Q. Guo, L. Wang, S. Zhang, and T. Chen. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4850–4861, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671576. URL https://doi.org/10.1145/3637528.3671576.

[7] Z. Chen, A. H. Cano, A. Romanou, and et al. MEDITRON-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[8] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[9] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

[10] DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

[11] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[12] H. Desai, K. Narasimhan, T. Liu, and et al. Medexqa: A challenging benchmark for long-form medical question answering. *arXiv preprint arXiv:2406.06331*, 2024.

[13] A. I. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1):50–79, 2009. ISSN 0376-0421. doi: https://doi.org/10.1016/j.paerosci.2008.11.001. URL https://www.sciencedirect.com/science/article/pii/S0376042108000766.

[14] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

[15] L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

[16] Z. Guo et al. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.

[17] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[18] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[19] M. A. Islam and M. M. Poly. Reinforcement learning in personalized medicine: A comprehensive review of treatment optimization strategies. *Cureus*, 17(4):e75123, 2025. doi: 10.7759/cureus.75123.

[20] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL https://arxiv.org/abs/2009.13081.

[21] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[22] N. Khandekar, Q. Jin, G. Xiong, S. Dunn, S. S. Applebaum, Z. Anwar, M. Sarfo-Gyamfi, C. W. Safranek, A. A. Anwar, A. Zhang, A. Gilson, M. B. Singer, A. Dave, A. Taylor, A. Zhang, Q. Chen, and Z. Lu. Medcalc-bench: Evaluating large language models for medical calculations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 84730–84745. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/99e81750f3fdfcaf9613db2dbf4bd623-Paper-Datasets_and_Benchmarks_Track.pdf.

[23] Y. Labrak, A. Bazoge, E. Morin, and et al. Biomistral: Open-source pre-trained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[24] J. Lee, W. Yoon, S. Kim, and et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

[25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

[26] S. Li, V. Balachandran, S. Feng, J. Ilgen, E. Pierson, P. W. W. Koh, and Y. Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

[27] Z. Li, Z. Lu, Y. Luo, and et al. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10 (7155), 2020. doi: 10.1038/s41598-020-62922-y.

[28] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, et al. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*, 2023.

[29] Z. Luo, Y. Song, Y. Gu, and et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *arXiv preprint arXiv:2210.10341*, 2022.

[30] T. Olatunji, C. Nimo, A. Owodunni, T. Abdullahi, E. Ayodele, M. Sanni, C. Aka, F. Omofoye, F. Yuehgoh, T. Faniran, et al. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*, 2024.

[31] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[32] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

[33] J. Peng, B. Yan, C. Raffel, and et al. Biomedlm: A 2.7b-parameter language model trained on pubmed abstracts and full articles. *arXiv preprint arXiv:2403.18421*, 2024.

[34] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

[35] A. U. Rahman, F. Al-Obeidat, A. Tubaishat, B. Shah, and S. Anwar. Data quality, bias, and strategic challenges in reinforcement learning for healthcare: A survey. 2024.

[36] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, and A. M. Rush. Multitask prompted training enables zero-shot task generalization, 2022. URL https://arxiv.org/abs/2110.08207.

[37] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM, Mar. 2025. doi: 10.1145/3689031.3696075.

[38] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. Large language models encode clinical knowledge, 2022. URL https://arxiv.org/abs/2212.13138.

[39] K. Singhal, T. Tu, J. Gottweis, and et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[40] G. Wang, J. Ran, R. Tang, C.-Y. Chang, Y.-N. Chuang, Z. Liu, V. Braverman, Z. Liu, and X. Hu. Assessing and enhancing large language models in rare disease question-answering. *arXiv preprint arXiv:2408.08422*, 2024.

[41] X. Wang, G. H. Chen, D. Song, Z. Zhang, Z. Chen, Q. Xiao, F. Jiang, J. Li, X. Wan, B. Wang, and H. Li. Cmb: A comprehensive medical benchmark in chinese, 2024. URL https://arxiv.org/abs/2308.08833.

[42] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, M. Patel, K. K. Pal, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. K. Sampat, S. Doshi, S. Mishra, S. Reddy, S. Patro, T. Dixit, X. Shen, C. Baral, Y. Choi, N. A. Smith, H. Hajishirzi, and D. Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022. URL https://arxiv.org/abs/2204.07705.

[43] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023. URL https://arxiv.org/abs/2212.10560.

[44] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[45] C. Wu, W. Lin, X. Zhang, and et al. PMC-LLaMA: Towards building open-source language models for medicine. *arXiv preprint arXiv:2304.14454*, 2023.

[46] C. Wu, P. Qiu, J. Liu, and et al. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8(58), 2025. doi: 10.1038/s41746-024-01390-4.

[47] Q. Xie, Q. Chen, A. Chen, and et al. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8(141), 2025. doi: 10.1038/s41746-025-01533-1.

[48] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, and et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[49] X. Yang, A. Chen, N. PourNejatian, and et al. GatorTron: A large clinical language model to unlock patient information from unstructured EHRs. *arXiv preprint arXiv:2203.03540*, 2022.

[50] Y. Zhu, Z. Huang, L. Mu, Y. Huang, W. Nie, S. Zhang, P. Liu, and X. Zhang. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*, 2025.

[51] Y. Zuo, S. Qu, Y. Li, Z. Chen, X. Zhu, E. Hua, K. Zhang, N. Ding, and B. Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.