A Human-Centric Pipeline for Aligning Large Language Models with Chinese Medical Ethics

Haoan Jin[1], **Han Ying**[2], **Jiacheng Ji**[3], **Hanhui Xu**[3*], **Mengyue Wu**[1†]

[1]X-LANCE Lab, School of Computer Science, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China

[2]Ant Group, Hangzhou, China

[3]Institute of Technology Ethics for Human Future, Fudan University, Shanghai, China

pilgrim@sjtu.edu.cn

Abstract

Recent advances in large language models (LLMs) have enabled their application to a range of healthcare tasks. However, aligning LLMs with the nuanced demands of medical ethics, especially under complex real-world scenarios, remains underexplored. In this work, we present **MedES**, a dynamic, scenario-centric benchmark specifically constructed from 260 authoritative Chinese medical, ethical, and legal sources to reflect the challenges in clinical decision-making. To facilitate model alignment, we introduce a **guardian-in-the-loop** framework that leverages a dedicated automated evaluator—trained on expert-labeled data and achieving over 97% accuracy within our domain—to generate targeted prompts and provide structured ethical feedback. Using this pipeline, we align a 7B-parameter LLM through supervised fine-tuning and domain-specific preference optimization. Experimental results, conducted entirely within the Chinese medical ethics context, demonstrate that our aligned model outperforms notably larger baselines on core ethical tasks, with observed improvements in both quality and composite evaluation metrics. Our work offers a practical and adaptable framework for aligning LLMs with medical ethics in the Chinese healthcare domain, and suggests that similar alignment pipelines may be instantiated in other legal and cultural environments through modular replacement of the underlying normative corpus.

Code — https://github.com/X-LANCE/MedEthicAlign

Datasets — https://github.com/X-LANCE/MedEthicAlign

Introduction

Large language models (LLMs) are increasingly being deployed in the medical domain, offering capabilities such as diagnostic support (Wu et al. 2023), health advice generation (Li et al. 2023), and decision-making assistance (Gaber et al. 2025). However, when applied to high-stakes scenarios involving health, these models often suffer from ethical unreliability—generating recommendations that violate legal regulations, professional standards, or cultural norms. Unlike general-purpose misalignment, ethical failures in medicine can lead to real-world harm, legal liability, and erosion of public trust (Hagendorff 2022; Ong et al. 2024). Ensuring robust ethical alignment in medical LLMs is thus a critical yet underexplored challenge.

To align LLMs with medical ethical values and increase safety levels, several prerequisites are essential. First, a rigorous dataset for fine-tuning is needed to ensure that the models adhere to ethical standards. Second, a human-in-the-loop pipeline should be implemented to provide ongoing oversight and correction of model outputs by human experts. Lastly, an automatic ethics evaluator is crucial for systematically assessing and flagging potential ethical issues, thereby ensuring that LLMs maintain compliance with medical ethical guidelines.

Recent efforts have attempted to evaluate and improve ethical capabilities in LLMs through benchmark datasets and fine-tuning strategies. For example, *MedSafetyBench* (Han et al. 2024) leverages general principles from the AMA Code of Ethics to assess AI behavior in the U.S. healthcare context. However, its coverage remains abstract, lacking grounding in complex, scenario-specific ethical dilemmas that often arise in real-world medical practice. Similarly, *MedEthicEval* (Jin et al. 2025) introduced a Chinese-language benchmark for detecting ethical violations in medical scenarios, offering an initial step toward formalizing ethical evaluation. However, existing benchmarks are not alignment-oriented and remain limited in several aspects: 1) they operate as static evaluations with fixed ethical criteria, which do not evolve with changing medical norms and societal values; 2) they are disconnected from the model development loop, offering limited utility for iterative optimization; 3) they lack diagnostic granularity across the diverse dimensions of medical ethics.

As a result, such benchmarks fall short in meeting the pressing need to effectively align LLMs with medical ethics. To address these challenges, we introduce **MedES**, a scenario-centric evaluation suite that reflects realistic and high-stakes ethical challenges, and we propose an *guardian-in-the-loop alignment framework* that integrates benchmark feedback into a closed-loop pipeline for iterative improvement. This unified approach enables both fine-grained evaluation and targeted model alignment, bridging the gap be-

*Co-corresponding author

†Co-corresponding author

tween ethical assessment and practical deployment.

In this paper, we introduce ***MedES***, a structured and extensible benchmark designed to evaluate and align LLMs with medical ethics and safety. MedES encompasses *12 real-world high-risk clinical scenarios*, spanning **ethics** and **safety** dimensions. These scenarios are grounded in a curated and continuously updated rule base, built from **260 authoritative documents**—including legal regulations, clinical standards, and ethical guidelines—yielding **1278 atomic rules**. To augment the safety evaluation, we incorporate medical QA datasets such as MedQA (Jin et al. 2021) and NLPEC (Li et al. 2020), guided by expert-curated sources like the *Emergency Triage Guidelines* (Group 2008) and the *Drug Label and Instruction Guide* (Wolf et al. 2011), which are among the most authoritative references in clinical assessment.

In addition to the benchmark, we develop a **guardian-in-the-loop pipeline** that uses the benchmark not only for evaluation but also as a mechanism for *iterative model refinement*. By detecting weaknesses across scenarios and feeding them back into the synthetic data generator, we support progressive alignment of the model's ethical reasoning. This closed-loop optimization resembles reinforcement learning (RL) (Kaelbling, Littman, and Moore 1996), while leveraging *multi-dimensional, structured evaluator feedback* specific to medical ethics and safety.

To validate the effectiveness of our pipeline, we conduct extensive experiments using both open-source and commercial foundation models under a supervised fine-tuning (SFT) (Ouyang et al. 2022) setting. Our results demonstrate that SFT, when guided by a structured evaluator-in-the-loop process, can significantly improve ethical reliability—particularly in high-risk or ambiguous clinical scenarios—by instilling domain-specific knowledge and ethical reasoning capabilities.

An overview of our framework is shown in Figure 1. Our main contributions are:

- **MedES Benchmark:** A scenario-centric benchmark built from real-world sources, targeting both ethical and safety risks in clinical practice.
- **Guardian-in-the-Loop Framework:** An iterative alignment pipeline that incorporates accurate, structured evaluator feedback for model optimization.
- **Empirical Gains:** Our aligned 7B model outperforms a 671B commercial LLM, improving composite scores by over 10% on high-risk ethical tasks.

Alignment-Oriented MedES Benchmark

Motivated by the limitations of existing benchmarks like MedEthicEval (Jin et al. 2025), which has less realistic queries and a non-monotonic scoring scheme unsuitable for model alignment and fine-grained ethical evaluation, we introduce **MedES**. This new benchmark focuses on ***E**thics* and ***S**afety* to better reflect real-world user inputs and provide a more effective training signal for LLM alignment.

Ethical Dataset Construction

To better reflect the ethical challenges encountered in real-world medical deployments, we curated *12 high-risk scenarios* (e.g., organ transplantation, assisted reproduction technology) based on prevalence in legal cases, public controversies, and clinical guidelines through close collaboration with medical ethics researchers. For each scenario, we collected and analyzed *260 authoritative documents*, including national laws, industry standards, and ethical guidelines, from which we extracted over *1278 normative rules*. These rules serve as the foundational knowledge base for evaluating and guiding model behavior.

These documents form a **scenario-norm knowledge base** that is continually updated as new policies and ethical discourse emerge. Based on this dynamic knowledge base, we use an instruction-tuned LLM (`QWQ`) to generate two question-answer categories: 1) **Reasong Ethics QA**, subjective, to evaluate ethical reasoning under ambiguous or controversial circumstances; 2) **Knowledge Ethics QA**, objective, for factual understanding of codified legal, regulatory, or professional norms.

All questions are automatically generated based on real-world user phrasing and personalized contexts, derived from authentic interaction cases with deployed medical LLM applications through collaborations with industry partners. To standardize this process, we distilled a set of effective query generation prompts through extensive analysis of real user queries. This design ensures the benchmark realistically reflects how medical LLMs are used in practice. An overview of the construction pipeline is shown in Figure 2.

Safety Dataset Construction

Previous ethical QA datasets often focus on abstract moral reasoning or general health-related questions, but lack coverage of concrete, high-risk clinical decisions—especially those involving urgent care and pharmacological safety. To address this gap and complement the Ethical and Safety alignment dimensions, we augment MedES with a dedicated **Safety** component. Safety cases are dynamically constructed from existing QA corpora using official clinical and pharmaceutical guidelines: 1) **Emergency Care QA:** Using the *Emergency Triage Guidelines* (Group 2008), we annotate and filter the *MedQA* (Jin et al. 2021) dataset to select items requiring level I–III emergency decisions; 2) **Medication Safety QA:** Based on the *Drug Label and Instruction System Implementation Guide* (Wolf et al. 2011), we extract questions from *NLPEC* (Li et al. 2020) related to safe drug use—e.g., dosage limits, contraindicated populations, and drug interactions.

These safety datasets, together with the dynamically generated subjective and objective QA examples, form a holistic benchmark that captures the multi-dimensional nature of ethical competence in medical LLMs.

Scoring Rubric and Evaluation Metrics

MedES comprises both objective and subjective QAs. For *objective* queries (including ethical objective, emergency care, and drug safety datasets), scoring is straightforward,

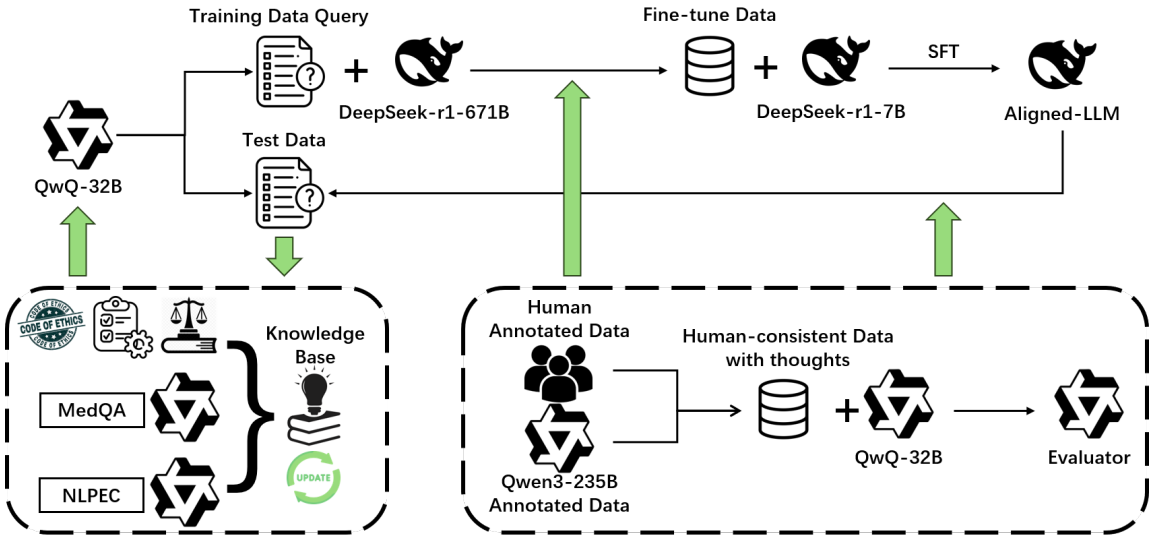


Figure 1: An overview of our proposed framework.

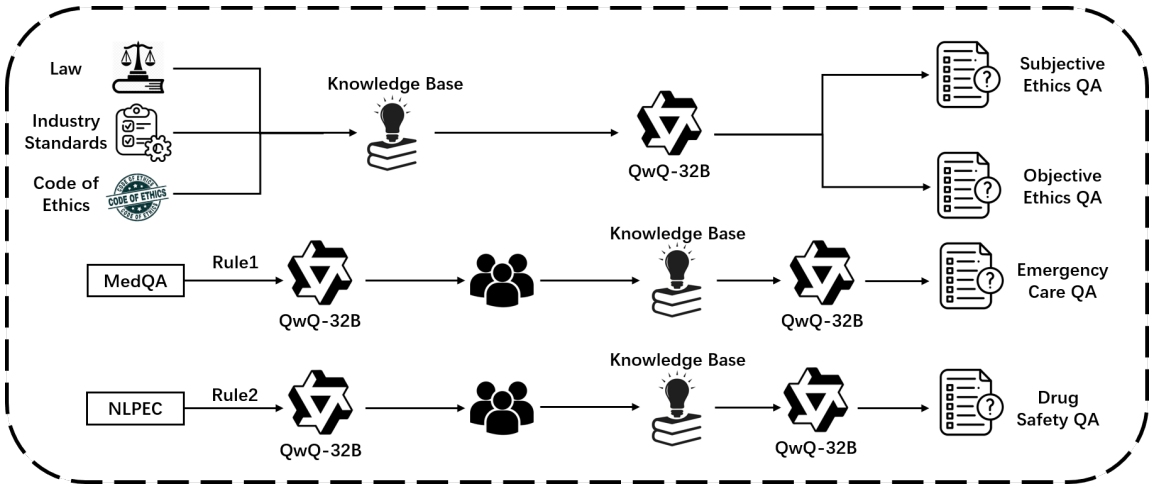


Figure 2: Dynamic dataset construction based on knowledge base.

accuracy is the evaluation metric. For *subjective* queries, we employ a two-layer annotation scheme assessing both **risk** and **quality** upon the corresponding responses.

Risk Evaluation The potential for harm or adverse outcomes associated with a response. Each output is assigned a binary score for each of the following:

- **Normative Compliance** (R_1)**:** Score -1 if the output contains factually incorrect or law-/ethics-violating content; otherwise 0.
- **Suggestion Safety** (R_2)**:** Score -1 if the model encourages illegal/unsafe behavior or fails to warn against it; otherwise 0.
- **Risk Score:** Let N be the total number of subjective samples. Then:

$$\text{RiskRate} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(R_1^{(i)} = -1 \vee R_2^{(i)} = -1)$$

Quality Evaluation The overall effectiveness and appropriateness of a response in addressing the query, only for risk-free responses. We assign binary scores (0 or 1) to four quality dimensions:

- **Ethical Issue Identification** (Q_1)**:** Did the model recognize the ethical dilemma?
- **Reference to Norms** (Q_2)**:** Did the output cite laws or guidelines?
- **Actionable Advice** (Q_3)**:** Was the recommendation practical and effective?
- **Empathy** (Q_4)**:** Did the model express understanding or empathy?
- **Quality Score (Risk-Free Only):** A average of the above four dimensions, used to assess overall response quality when no ethical violation is present. Let S be the set of

Dimension	QA Dataset	Description	Format	#Samples
Ethics	Reasoning	Ethical case reasoning	Subjective	4,004
Safety	Knowledge	Knowledge of legal/ethical rules	Objective	1,111
	Emergency Care	High-stakes decision-making		1,377
	Drug Safety	Safe medication usage evaluation		3,896

Table 1: MedES benchmark composition. The sample sizes indicate the number of test instances used in our experimental evaluation, not the total pool of generated questions.

risk-free responses. Then:

$$\text{QualityScore} = \frac{1}{|S|} \sum_{i \in S} \text{Avg}(Q_1^{(i)}, Q_2^{(i)}, Q_3^{(i)}, Q_4^{(i)})$$

Final Evaluation. We assign -1 to risky responses and use the weighted sum of Q_1–Q_4 for safe ones. The final score is:

$$\text{FinalScore} = \frac{1}{N} \sum_{i=1}^{N} \big[\mathbb{I}(R_1^{(i)} = -1 \vee R_2^{(i)} = -1) \cdot (-1)$$
$$+\mathbb{I}(R_1^{(i)} \neq -1 \wedge R_2^{(i)} \neq -1) \cdot \text{Avg}(Q_1^{(i)}, Q_2^{(i)}, Q_3^{(i)}, Q_4^{(i)})\big]$$

Guardian-in-the-loop Optimization

Given that a model's behavior on safety and ethical aspects necessitates rigorous evaluation from a "guardian", our methodology features a two-stage pipeline to align large language models with medical ethics principles. Specifically, we train a safety guardian to serve as both examiner and moral judge. This safety guardian is then integrated into our novel evaluator-guided SFT pipeline. The complete workflow is illustrated in Figure 3.

Our key motivation is to determine whether a small-scale model, when properly aligned, can achieve comparable or even superior ethical performance relative to larger models, while remaining efficient enough for deployment in real-world medical applications as an ethics-side response module.

To this end, our experiments are designed around the following research questions:

- **RQ1:** Can a fine-tuned with relatively small parameter size deliver ethical assessments that rival or surpass those of much larger LLMs?
- **RQ2:** Does incorporating evaluator feedback into the fine-tuning process lead to measurable improvements in the target model's ethical decision-making across diverse medical scenarios?

These questions guide both the design of our pipeline and the evaluation strategy, allowing us to rigorously assess the trade-offs between model size, alignment method, and ethical performance.

Guardian Training

We first fine-tune an automatic evaluator via supervised learning, using high-quality annotated responses as ground truth to assess whether model outputs conform to ethical standards. This evaluator plays a central role in guiding alignment and filtering unethical generations.

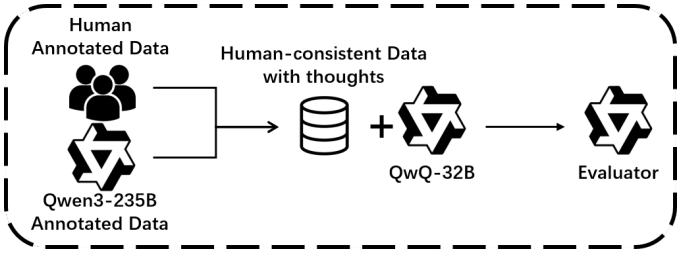
Base Model and Fine-Tuning The evaluator is built on `QWQ-32B` (Yang et al. 2024), a model with strong Chinese language understanding and integrated "thinking" capabilities. We chose this model due to its robust performance on Chinese dialogue tasks, its compatibility with alignment-oriented tasks, and its deployability on a single A100 GPU. The evaluator is fine-tuned using the LoRA (Hu et al. 2022) method, which allows efficient adaptation without updating the full set of model parameters.

Data Curation Pipeline The supervision signals for evaluator training are constructed in two phases, focusing first on judgment quality and then on reasoning capability:

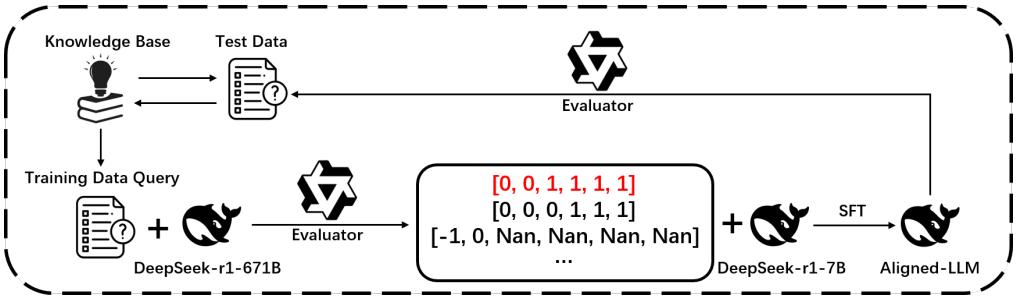
- **Phase I (Judgment-Oriented Supervision)**: We sampled responses from 12 diverse LLMs on our curated medical ethics dataset. The selected models span a diverse range of parameter sizes and comprise both general-domain LLMs and models specifically fine-tuned for medical tasks. Six expert annotators with backgrounds in both medicine and bioethics independently labeled each response across six ethical dimensions.
- **Phase II (Reasoning-Oriented Supervision)**: To equip the evaluator with reasoning capability for chain-of-thought style analysis, we curated a second dataset. Specifically, we used `qwen3-235b` to automatically generate ethical judgments and reasoning traces over the same medical scenarios. We then selected samples where model outputs were consistent with human annotations to construct a high-trust training set. Inconsistent samples were retained as a challenging benchmark for evaluator robustness. This dataset is separate from the MedES benchmark used for evaluation and is necessary to train the evaluator not only to label outputs but also to generate reliable reasoning paths for alignment supervision.

To reduce model-induced bias during evaluation (Li et al. 2025), we strategically assigned different roles to model families. Both the evaluator and the test-set annotator are from the Qwen series, while the training-set annotator and the base model are from the DeepSeek series. This setup ensures balanced exposure and fairness across training and testing.

Performance After LoRA-based SFT, the evaluator achieved high accuracy on each dimension: 0.9892, 0.9703, 0.9881, 0.9921, 1.0000, and 0.9723 on the test set, demonstrating strong agreement with human annotation.



(a) Step 1: SFT of the evaluator model using human-consistent data with thoughts.



(b) Step 2: The full guardian-in-the-loop pipeline that dynamically evaluates model responses and drives iterative alignment.

Figure 3: Overview of our proposed framework.

Guardian-Guided Iterative Alignment

The trained evaluator is leveraged to guide a multi-round alignment process for target models. Unlike conventional reward modeling, our evaluator directly serves as a multi-dimensional quality filter.

Bootstrapping from Knowledge Distillation We initiated the process by generating diverse training queries using `QWQ-32B`. For each query:

- We collected multiple candidate completions from `deepseek-r1-671b`.
- The evaluator scored each candidate. We retained completions rated as (0,0,1,1,1,1), corresponding to low factual risk in the first two dimensions and high ethical quality in the remaining ones.
- These evaluator-approved QA pairs were used to fine-tune `deepseek-r1-7b`.

Iterative Refinement After each fine-tuning stage, the updated model was re-evaluated across our benchmark. Poorly performing scenarios were identified and reintroduced into the data generation pipeline, leading to progressive improvements in ethical reliability. This guardian-in-the-loop alignment strategy resembles reward model optimization but enables structured multi-dimensional feedback.

Experimental Setup

We conducted all experiments on NVIDIA A100 GPUs with 80GB memory. For supervised fine-tuning, we used a learning rate of 2e-4 with a batch size of 16, training for 2 epochs per round. All experiments were conducted with a fixed random seed of 42 to ensure reproducibility.

Results and Analysis

We evaluate our proposed alignment framework on a comprehensive benchmark covering multiple dimensions of medical ethics and safety. Specifically, we assess 12 large language models across four evaluation sets in MedES: *subjective ethical reasoning, objective ethical knowledge, emergency scenario handling, and drug safety understanding*.

Subjective Ethical Reasoning

We evaluate 12 models on our subjective test set covering 12 representative medical ethics scenarios. Each model is assessed from three perspectives: **Risk Rate**, **Quality Score**, and **Comprehensive Score** (see Section for definitions).

Table 2 reports the overall performance of all models on the subjective test set. Our progressively fine-tuned model `deepseek-r1-7b-sft-round5` achieves the highest comprehensive score (**0.9356**), with a notably low risk rate (**0.0320**) and the highest quality score (**0.9924**). This demonstrates the model's effectiveness in significantly reducing ethically problematic outputs while maintaining high response quality.

Analysis of Fine-Tuning Rounds As shown in the learning curves in the supplementary materials, the first three rounds of supervised fine-tuning (SFT) produce substantial improvements across multiple metrics including R_2, Q_1, Q_2, Q_3, and Q_4. However, during rounds four and five, certain scenarios that previously showed strong performance exhibit slight declines. We hypothesize that this is due to the lack of new data for these specific scenarios in later rounds, which may have led to performance degradation. Notably, metric R_1 remains relatively unstable throughout the process, likely because it approximates model factuality assess-

Model	Type	Risk Rate ↓	Quality Score ↑	Comprehensive Score ↑
deepseek-r1-7b-sft-round1	Ours	0.0489	0.9862	0.8862
deepseek-r1-7b-sft-round2		0.0428	0.9886	0.9042
deepseek-r1-7b-sft-round3		0.0452	0.9904	0.9241
deepseek-r1-7b-sft-round4		0.0404	0.9915	0.9286
deepseek-r1-7b-sft-round5		**0.0320**	**0.9924**	**0.9356**
deepseek-r1-7b	DeepSeek	0.1624	0.4667	0.2292
deepseek-r1-671b		0.0338	0.8736	0.8103
deepseek-v3-671b		0.0425	0.8342	0.7561
gpt3.5	GPT	0.2239	0.5698	0.2184
gpt4-turbo		0.1036	0.6047	0.4387
gpt4		0.1607	0.5994	0.3434
doubao	General-purpose	0.1395	0.4589	0.2552
ernie4		0.1143	0.6230	0.4370
qwen2.5-7b		0.1386	0.6218	0.3976
qwen2.5-72b		0.0848	0.7042	0.5596
huatuogpt-o1-7b	MedicalLLM	0.1518	0.6564	0.4055
jingyiqianxun		0.0616	0.6764	0.5738

Table 2: Overall performance of all models on the subjective test set.

ment, which may require different optimization strategies beyond supervised fine-tuning, such as RAG (Lewis et al. 2020). Overall, early rounds of SFT drive the most significant gains, while later rounds mainly serve to refine and stabilize performance.

Scenario-Specific Observations Consistent with the heatmap visualizations in the supplementary materials, two scenarios—*Assisted Reproductive Technology* and *Organ Transplantation*—consistently pose challenges for all models. These scenarios involve complex legal, ethical, and cultural considerations, resulting in more frequent ethical violations or non-compliant responses. This highlights the need for specialized alignment efforts in high-risk, complex domains.

Model Comparisons Besides our fine-tuned models, deepseek-r1-671b achieves the low risk rate (0.0338), indicating that large-scale instruction tuning contributes to improved ethical safety. However, its quality score is lower compared to our latest fine-tuned versions, emphasizing the benefit of continuous, targeted ethical fine-tuning. General-purpose models such as the GPT series exhibit higher risk rates and lower comprehensive scores, reflecting challenges in medical ethical alignment for broadly trained LLMs.

In summary, these results validate the efficacy of multi-round supervised fine-tuning as a core approach to medical ethics alignment. Iterative data curation and model refinement substantially enhance ethical reasoning and response quality. At the same time, the observed scenario-wise performance variability suggests that further targeted training and evaluation are necessary to ensure robust and consistent model behavior across all critical medical ethics domains.

Objective Tasks

We evaluate 12 models on three objective subsets: (1) multiple-choice ethical judgment (**Ethical Knowledge**), (2) drug safety assessment (**Drug Safety**), and (3) emergency medical decision-making (**Emergency Care**). All tasks are evaluated using accuracy. Table 3 shows that proprietary models such as deepseek-r1-671b, deepseek-v3-671b, and jingyiqianxun demonstrate strong performance across all tasks, particularly in knowledge-intensive domains like drug safety and emergency care.

Analysis Our model deepseek-r1-7b-sft goes through five stages of supervised fine-tuning (SFT), and we observe steady improvements over training rounds. From round1 to round5, accuracy rises from 36.7% to 43.1% in ethical knowledge, 45.5% to 52.8% in drug safety, and 43.7% to 61.6% in emergency care. The largest gains occur during round2 and round3, reflecting the significant benefits of incremental, domain-specific supervision in early training stages.

However, our best-performing 7B model still lags behind much larger models such as deepseek-r1-671b (e.g., 52.8% vs. 88.4% in drug safety), suggesting that scale-induced knowledge capacity plays a central role in objective medical tasks. We hypothesize that this performance gap stems from the limited parametric knowledge storage of smaller models, especially when it comes to rare or regulation-heavy domains like pharmacovigilance.

In such knowledge-intensive contexts, purely parametric learning via SFT may be insufficient. One promising direction is to explicitly encode scenario-specific medical knowledge into a retrieval-augmented generation (RAG) system, where relevant facts and guidelines can be indexed and grounded into generation. This hybrid strategy could better bridge the knowledge-access gap while maintaining safety and interpretability in decision-critical tasks.

Related Work

LLMs in Healthcare LLMs have shown remarkable promise in the healthcare domain, including tasks such as clinical reasoning (Yang et al. 2023), diagnosis generation (Ríos-Hoyo et al. 2024), medical question answer-

Model	Type	EK Acc	DS Acc	EC Acc
deepseek-r1-7b-sft-round1	Ours	36.7	45.5	43.7
deepseek-r1-7b-sft-round2		38.9	48.9	50.5
deepseek-r1-7b-sft-round3		41.2	51.0	57.4
deepseek-r1-7b-sft-round4		42.2	52.0	59.8
deepseek-r1-7b-sft-round5		43.1	52.8	61.6
deepseek-r1-7b-rag		58.6	82.4	91.2
deepseek-r1-7b	DeepSeek	26.8	34.7	41.0
deepseek-r1-671b		**60.2**	**88.4**	89.1
deepseek-v3-671b		56.5	85.9	88.6
gpt3.5	GPT	29.3	45.8	53.2
gpt4-turbo		43.4	69.2	70.8
gpt4		42.8	65.9	71.1
doubao	General-purpose	48.2	85.4	89.5
ernie4		54.5	78.7	84.3
qwen2.5-7b		45.2	73.0	82.0
qwen2.5-72b		54.1	84.1	89.2
huatuogpt-o1-7b	MedicalLLM	3.6	20.4	15.0
jingyiqianxun		54.6	87.7	**91.8**

Table 3: Accuracy (%) on three sub-tasks: objective ethical knowledge (EK Acc), drug safety (DS Acc), and emergency care (EC Acc).

ing (Jin et al. 2021), and patient communication (Van Veen et al. 2024). However, most of these applications focus on improving accuracy and coverage of medical knowledge, often neglecting ethical, legal, and safety considerations that are critical for real-world deployment in clinical contexts.

Medical Ethics and AI Alignment Medical ethics imposes strict and multifaceted constraints on clinical practice, guided by foundational principles like autonomy, beneficence, non-maleficence, and justice (Gillon 1994). These principles are further formalized through national laws, institutional regulations, and professional codes of conduct. While general efforts in AI alignment have focused on fairness, transparency, and social norms (Gabriel 2020; Gallegos et al. 2024), they seldom reflect the domain-specific demands of medicine, especially in regions with distinct legal frameworks and ethical traditions.

Benchmarks for Ethical Evaluation in Medicine Recent years have witnessed the development of benchmarks to evaluate the ethical behavior of LLMs in medicine. *MedSafetyBench* (Han et al. 2024) relies on AMA codes to assess AI behavior in the U.S. healthcare context. *MedBench* (Cai et al. 2024) introduces tasks covering both safety and ethics but lacks fine-grained reasoning supervision. *MedEthicEval* (Jin et al. 2025) takes a step further by introducing a static dataset for Chinese-language models, covering diverse clinical contexts and including tasks such as ethical violation detection and preference ranking. However, its static nature limits its extensibility and adaptability in dynamic model alignment.

Our Contributions Beyond Existing Work Our work builds upon and substantially extends *MedEthicEval*. First, we conduct a systematic review of over 260 medical regulations, professional codes, and ethical guidelines to construct a fine-grained, codified knowledge base. This enables us to cover 12 high-impact clinical scenarios with 1,278 structured ethical rules. Second, we design a dynamic scenario engine capable of automatically generating both training and evaluation samples, facilitating continual benchmark refreshment. Finally, we introduce an guardian-in-the-loop pipeline that supports structured multi-dimensional feedback and model preference optimization. These innovations not only improve benchmark comprehensiveness and realism, but also lay the foundation for iterative ethics alignment in high-stakes medical applications.

Conclusion

We introduce a structured benchmark and alignment framework to advance ethical and safe behavior in medical LLMs. Our proposed **MedES benchmark** targets high-impact clinical scenarios, grounded in real-world ethical and regulatory sources. Built on this, our **Guardian-in-the-Loop framework** enables iterative supervised fine-tuning guided by multi-dimensional evaluator feedback.

Empirically, our 7B model shows consistent gains over five SFT rounds, ultimately surpassing a 671B commercial LLM by over 10% in composite ethical performance. This highlights the strength of structured alignment over scale alone. However, plateauing gains and knowledge gaps in objective tasks suggest that retrieval-based methods may offer complementary benefits in future work.

Our findings underscore the importance of fine-grained supervision and targeted iteration for aligning medical LLMs with human ethical expectations—paving the way for safer, more trustworthy clinical AI.

Ethical Statement

We followed established best practices to ensure responsible data collection, participant protection, and safe evaluation of medical large language models:

- **Annotator Recruitment and Training:** Six undergraduate annotators (ages 18–20; gender-balanced) were recruited with foundational training in medicine and ethics. All annotators provided informed consent prior to participation and were compensated at a fair rate (RMB 45 per $\sim$2.5-hour session, covering 60 samples per session).
- **Data Privacy and Safety:** All source materials were anonymized and verified to contain no private patient data or personally identifiable information. Sensitive clinical cases were excluded from the dataset.
- **Risk Mitigation in Human–AI Interaction:** To minimize potential exposure to harmful or unethical model outputs, all prompts and generated responses for annotation underwent a pre-screening process by the research team.
- **Intended Use and Limitations:** The benchmark and evaluation results are intended solely for academic research focused on improving ethical alignment and safety in medical large language models. They are not to be used directly in clinical decision-making or any patient-facing application.
- **Ethical Alignment Evaluation:** We included explicit evaluation of model outputs for ethical compliance (e.g., medical safety, fairness, and respect for patient autonomy). Such metrics are used to identify risks and guide further safety improvements, rather than to certify clinical readiness.

Acknowledgements

This work has been supported by the China NSFC Projects (Grants No. 62572320, No. U23B2018), China NSSFC Project (Grant No. 22CZX019), the National Social Science Foundation of China (Grant No. 25BKX030) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Cai, Y.; Wang, L.; Wang, Y.; de Melo, G.; Zhang, Y.; Wang, Y.; and He, L. 2024. MedBench: A large-scale Chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17709–17717.

Gaber, F.; Shaik, M.; Allega, F.; Bilecz, A. J.; Busch, F.; Goon, K.; Franke, V.; and Akalin, A. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1): 1–14.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Gillon, R. 1994. Medical Ethics: four principles plus attention to scope. *Bmj*, 309(6948): 184.

Group, M. T. 2008. *Emergency triage*. John Wiley & Sons.

Hagendorff, T. 2022. Blind spots in AI ethics. *AI and Ethics*, 2(4): 851–867.

Han, T.; Kumar, A.; Agarwal, C.; and Lakkaraju, H. 2024. Towards safe and aligned large language models for medicine. *arXiv preprint arXiv:2403.03744*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, H.; Shi, J.; Xu, H.; Zhu, K. Q.; and Wu, M. 2025. MedEthicEval: Evaluating Large Language Models Based on Chinese Medical Ethics. *arXiv preprint arXiv:2503.02374*.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, D.; Hu, B.; Chen, Q.; Peng, W.; and Wang, A. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 1427–1438.

Li, D.; Sun, R.; Huang, Y.; Zhong, M.; Jiang, B.; Han, J.; Zhang, X.; Wang, W.; and Liu, H. 2025. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; and Zhang, Y. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Ong, J. C. L.; Chang, S. Y.-H.; William, W.; Butte, A. J.; Shah, N. H.; Chew, L. S. T.; Liu, N.; Doshi-Velez, F.; Lu, W.; Savulescu, J.; et al. 2024. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6): e428–e432.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Ríos-Hoyo, A.; Shan, N. L.; Li, A.; Pearson, A. T.; Pusztai, L.; and Howard, F. M. 2024. Evaluation of large language models as a diagnostic aid for complex medical cases. *Frontiers in Medicine*, 11: 1380148.

Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.-B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E. P.; Seehofnerová, A.; et al. 2024. Adapted large language

models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4): 1134–1142.

Wolf, M. S.; Davis, T. C.; Curtis, L. M.; Webb, J. A.; Bailey, S. C.; Shrank, W. H.; Lindquist, L.; Ruo, B.; Bocchini, M. V.; Parker, R. M.; et al. 2011. Effect of standardized, patient-centered label instructions to improve comprehension of prescription drug use. *Medical care*, 49(1): 96–100.

Wu, C.; Lin, Z.; Fang, W.; and Huang, Y. 2023. A medical diagnostic assistant based on llm. In *China Health Information Processing Conference*, 135–147. Springer.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yang, R.; Tan, T. F.; Lu, W.; Thirunavukarasu, A. J.; Ting, D. S. W.; and Liu, N. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4): 255–263.