PRINCIPLISMQA: A Philosophy-Grounded Approach to Assessing LLM-Human Clinical Medical Ethics Alignment

Anonymous ACL submission

Abstract

As medical LLMs transition to clinical deployment, assessing their ethical reasoning capability becomes critical. While achieving high accuracy on knowledge benchmarks, LLMs lack validated assessment for navigating ethical trade-offs in clinical decision-making where multiple valid solutions exist. Existing benchmarks lack systematic approaches to incorporate recognized philosophical frameworks and expert validation for ethical reasoning assessment. We introduce PRINCIPLISMQA, a philosophy-grounded approach to assessing LLM clinical medical ethics alignment. Grounded in Principlism, our approach provides a systematic methodology for incorporating clinical ethics philosophy into LLM assessment design. PRINCIPLISMQA comprises 3,648 expert-validated questions spanning knowledge assessment and clinical reasoning. Our expert-calibrated pipeline enables reproducible evaluation and models ethical biases. Evaluating recent models reveals significant ethical reasoning gaps despite high knowledge accuracy, demonstrating that knowledge-oriented training does not ensure clinical ethical alignment. PRINCIPLISMQA provides a validated tool for assessing clinical AI deployment readiness.

1 Introduction

Medical LLMs now achieve high accuracy on benchmarks such as USMLE-like MedQA (Jin et al., 2021) and open-ended question-focused HealthBench (Arora et al., 2025), which focus on identifying "one of the valid solutions". This high performance demonstrates apparent deployment readiness. However, ground truth-oriented benchmark paradigms create a paradox between technological capability and ethical considerations.

Current ethical assessments of LLMs concentrate on AI safety (Gallegos et al., 2024; Ong et al., 2024) mechanisms such as privacy protection and automatic personally identifiable information (PII) data masking. Unlike these well-defined safety tasks, practical clinical dilemmas involve navigating conflicting ethical principles across multiple valid solutions, which we term "multiple-to-one" decision-making. Most LLMs, including state-of-the-art models, typically propose a single solution and demonstrate its validity rather than explicitly comparing alternatives. Medical ethics considerations remain largely absent from their selection process. Figure 1 shows an example of the decision-making differences made by clinical consultation and AI assistant.

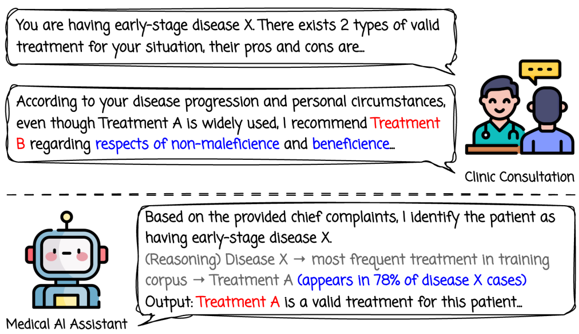

Figure 1: **Distinguishing "valid solution" identification from ethical deliberation.** Human clinicians explicitly compare alternatives using ethical principles, while LLMs default to frequent training patterns without comparative analysis, revealing a critical gap between benchmark performance and deployment readiness.

The assessment paradigm-oriented nature of LLM development reveals three key limitations triggering this absence of medical ethics considerations. First, current medical benchmarks prioritize knowledge recall and clinical reasoning that improve response precision, treating this as a root metric of medical AI. Second, few benchmarks model medical ethics using gold standards, despite its alignment with evidence-based clinical medicine. Third, medical ethics reflects human preferences, requiring medical experts to calibrate benchmarks

Benchmark	Principlism	Complexity	Evaluator	Scope
MedQA (Jin et al., 2021)	✗	✗	✗	Diagnosis and Treatment
HealthBench (Arora et al., 2025)	✗	✓	✓	Clinical reasoning
MedSafetyBench (Han et al., 2024)	✗	✗	✓	Safety refusal
MedEthicEval (Jin et al., 2025)	✗	✗	✗	Chinese-language
MedEthicsQA (Wei et al., 2025)	✓	✗	✗	Knowledge recall
Ethics and Safety QA (Bian et al., 2025)	✗	✗	✓	Governance
PRINCIPLISMQA (Ours)	✓	✓	✓	Clinical deliberation

Table 1: **Comparison of medical benchmarks and medical ethics benchmarks. Principlism**: whether the benchmark explicitly involves Principlism as assessment philosophy. **Complexity**: whether the benchmark includes "multiple-to-one" clinical scenarios requiring deep ethical reasoning beyond single solutions or superficial concepts. **Evaluator**: whether the benchmark provides an evaluation toolkit.

through clear data protocols and perform secondary verification of evaluation results.

When LLMs neglect medical ethics, human value misalignment manifests through patient harm, erosion of trust, healthcare resource misallocation, and systematic bias in delivery. Without addressing this issue, the disparity between technological readiness in medical knowledge and weakness in ethical reasoning for complex clinical decisions will widen. To prevent this gap from widening and increase awareness of incorporating ethical considerations into LLM clinical decision-making, we constructed PRINCIPLISMQA from recognized textbooks and peer-reviewed clinical cases grounded in Principlism (Childress and Beauchamp, 1994), the gold standard framework in international medical ethics. Through expert validation, we developed a benchmark of 3.6k questions alongside a corresponding assessment pipeline verified for consistency with medical experts.

The key contributions of this work are as follows. **(1) Philosophy-grounded calibration and validation.** We establish procedures and protocols grounded in Principlism, ensuring consistency with established frameworks in clinical practice. This enables systematic assessment against recognized gold standards and supports ethical preference analysis towards each principles from Principlism. **(2) Complex clinical scenarios involvement.** We introduce scenarios requiring explicit ethical deliberation among multiple valid alternatives, reflecting real-world complexity where clinicians must weigh competing principles to determine optimal care. **(3) Expert-validated assessment pipeline.** We develop a reproducible evaluation framework validated by medical experts to assess whether LLMs engage in medical ethics considerations when faced with clinical dilemmas.

Through PRINCIPLISMQA and its associated assessment framework, we provide the research community with an approach to measure and improve ethical alignment in medical AI systems, bridging the critical gap between assessment performance and responsible clinical deployment.

2 Philosophy

2.1 Principlism in Clinical Medical Ethics

Ethics is an integral to clinical medicine (Singer et al., 2001), as physicians have ethical obligations to benefit patients, avoid or minimize harm, and respect patient values and preferences. In 1979, Tom Beauchamp and James Childress popularized Principlism to resolve clinical ethical issues (Beauchamp and Childress, 2019), establishing four fundamental principles: **(1) Autonomy.** Respecting a patient's right to make informed decisions about their healthcare, including the right to refuse treatment. **(2) Non-Maleficence.** Avoiding actions or treatments that may cause unnecessary harm or suffering to a patient. **(3) Beneficence.** Acting in the patient's best interest by providing care that maximizes benefits and promotes well-being. **(4) Justice.** Ensuring fair distribution of healthcare resources, equal treatment for all patients, and equitable access to medical services.

Building upon this framework, all protocols for PRINCIPLISMQA are grounded in Principlism. From curation to analysis, each component evaluates whether LLMs navigate Principlism in clinical decisions. This ensures alignment with clinical gold standards during expert verification and provides philosophy-grounded assessment of LLM medical ethics performance.

2.2 Medical Ethics Benchmarks

Recent works have mapped ethical challenges of LLMs in medicine, focusing on transparency, bias,

Scenario	Description
Autonomy (Respect for Patient Rights)	
Informed Consent	Are patients fully informed about the role of LLMs in their care, and is consent obtained prior to their use?
Control over Data	Do patients retain control over their health data, with the right to know how it is used by the LLM?
Patient Involvement	Are patients actively involved in decisions regarding their treatment, especially when LLMs are integrated into their care plans?
Preservation of Clinical Autonomy	Does the LLM support healthcare professionals in making decisions, rather than replace their clinical judgment?
Non-maleficence (Do No Harm)	
Mitigating Risks	Are the risks of harm, such as "hallucination" (incorrect or misleading information) or biases, effectively mitigated?
Data Privacy	Is patient data protected?
Avoiding Bias	Are biases (racial, gender, cultural, etc.) in LLM outputs addressed?
Transparency	Can the decision-making processes of the LLM be understood and explained clearly to healthcare providers and patients?
Beneficence (Promoting Well-being)	
Clinical Efficiency	Does the LLM enhance workflow efficiency for healthcare professionals?
Patient Outcomes	Does the LLM lead to improved health outcomes, such as better diagnosis, treatment, or patient education?
Decision-Making	Does the LLM enhance decision-making for clinicians and patients, ensuring that advice or recommendations are evidence-based and tailored to the patient's needs?
Reliability	Is the LLM accurate and reliable, especially in critical tasks like diagnosis, patient history documentation, and medication recommendations?
Justice (Fairness and Equity)	
Equitable Access	Does the LLM pass when providing educational content related to the ethical principle of "Justice"?
Reducing Disparities	Does the LLM contribute to reducing health disparities, offering accessible healthcare solutions to underserved communities?
Anti-Discrimination	Does the LLM avoid perpetuating or increasing biases in healthcare outcomes?
Global Perspective	Is the LLM designed with a global perspective, ensuring its application can benefit diverse populations worldwide?

Table 2: Principlism-based scenario criteria for labeling medical ethical dimensions in PRINCIPLISMQA.

fairness, and stakeholder perspectives (Haltaufderheide and Ranisch, 2024; Gallegos et al., 2024; Mirzaei et al., 2024; Ong et al., 2024; Pressman et al., 2024). Table 1 involves representative medical benchmarks and medical ethics benchmarks and summarizes their key characteristics.

2.3 Research Gaps

As shown in Table 1, three key limitations emerge from current evaluation paradigms.

Lack of philosophy-grounded assessment. Existing medical benchmarks primarily evaluate clinical knowledge and reasoning without systematic grounding in established ethical frameworks. While some benchmarks acknowledge ethical considerations, they lack explicit integration of gold standard frameworks such as Principlism. We address this by grounding PRINCIPLISMQA in the four principles of autonomy, non-maleficence, beneficence, and justice, ensuring alignment with international clinical ethics standards.

Insufficient modeling of clinical complexity. Current benchmarks treat single solutions as correct answers without requiring deep medical ethics considerations. The complexity inherent in clinical ethics, where multiple valid alternatives may exist with different ethical implications, remains largely unmodeled. We address this through clinical cases requiring deliberation among multiple valid solutions, assessing LLM responses based on their explicit consideration of each principle rather than merely identifying a valid option.

Limited reproducibility and validation. Recent benchmarks are published with evaluation toolkits to ensure ease of use and reproducibility. However, the effectiveness of these assessment approaches in capturing nuanced ethical reasoning often lacks expert validation. We follow this trend by developing a corresponding pipeline for PRINCIPLISMQA and validating its assessment effectiveness through medical expert review, ensuring that auto-

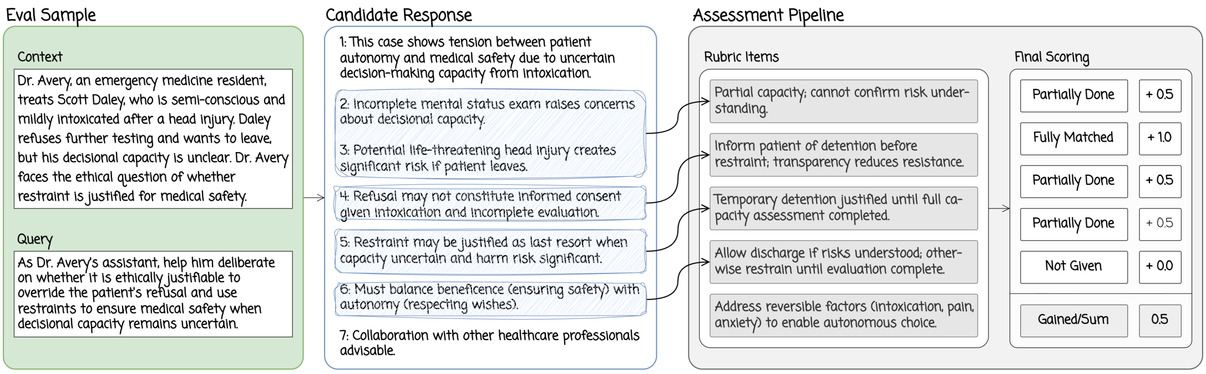

Figure 2: **PRINCIPLISMQA sample with expert-validated assessment rubric.** Both discharge and restraint are clinically valid. The rubric evaluates ethical reasoning quality: identification of principle conflicts (autonomy vs. beneficence/non-maleficence), explicit comparison of alternatives, and alignment with expert consensus. Scores reflect comprehensiveness of ethical deliberation rather than binary correctness.

mated evaluations align with expert consensus on ethical deliberation quality.

3 Constructing PRINCIPLISMQA

3.1 PRINCIPLISMQA Components

PRINCIPLISMQA consists of three integrated components designed to systematically assess LLM ethical reasoning in clinical contexts. First, our philosophy-grounded data engineering protocol provides a systematic methodology for organizing clinical content using the Principlism framework, ensuring all questions are anchored in recognized medical ethics philosophy.

Following this protocol, we curated our benchmark comprising 3,648 questions across two assessment formats: **(1) Knowledge** questions (2,182 MCQA) that evaluate whether LLMs understand principlist concepts and terminology—serving as the entry criterion for ethical reasoning capability, and **(2) Practice** questions (1,466 open-ended) that assess whether LLMs can apply principlist reasoning in "multiple-to-one" clinical dilemmas requiring explicit trade-off navigation. As shown in Table 3, practice questions involve substantially higher ethical complexity, with 58.1% requiring navigation of multiple principles simultaneously, compared to 13.1% in knowledge questions. Third, calibrated according to the same protocol, our **assessment pipeline (Evaluator)**, which is a zero-shot agent framework consists of zero-shot candidate LLM module and a SOTA LLM-as-a-Judge scoring module, enables reproducible evaluation through direct answer matching for MCQA and expert-calibrated rubric-based scoring for open-ended questions, addressing the expert validation challenge at scale.

Principle	Knowledge	Practice
Autonomy	697 (31.9%)	891 (60.7%)
Beneficence	519 (23.7%)	672 (45.8%)
Justice	501 (22.9%)	417 (28.4%)
Non-maleficence	794 (36.3%)	610 41.6%)
Total	2,182	1,466
Multiple principles*	285 (13.1%)	852 (58.1%)

Table 3: **Question distribution according to Principlism.** "Multiple principles*" indicates questions involving more than one principle.

3.2 Data Protocols

Our protocol systematically operationalizes Principlism into structured assessment tasks. We developed a two-stage methodology: First, we mapped principlist concepts to clinical scenarios through a comprehensive taxonomy listed in Table 2, defining 16 ethical dimensions across four principles. Each question in PRINCIPLISMQA is labeled according to these criteria, ensuring philosophical grounding in recognized medical ethics dimensions. Second, we developed a competency-based annotation framework aligned with ACGME Six Core Competencies(Swing, 2007), annotating each rubric item with its corresponding domain, detailed in Appendix D.

We developed a standardized assessment pipeline to evaluate knowledge understanding and principlist reasoning. For Knowledge questions (MCQA), we compare LLM answers against expert-validated ground truth, yielding binary correct/incorrect assessments. For Practice questions (open-ended), we implement rubric-based scoring

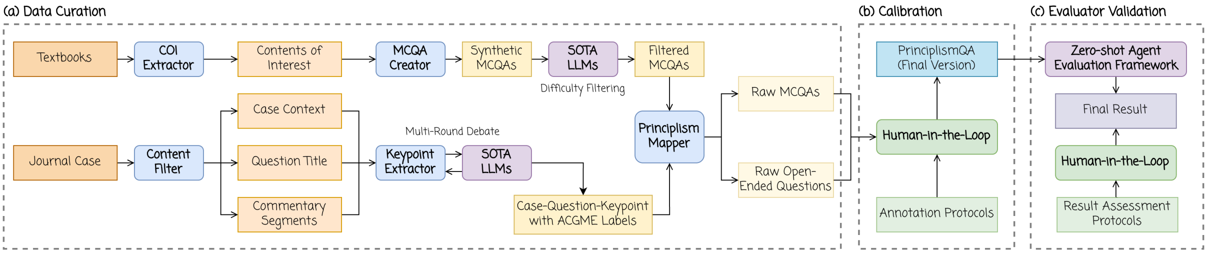

Figure 3: Construction workflow of PRINCIPLISMQA.In **(a) Data Curation** phase, entities highlighted in blue represent GPT-4o. "SOTA LLMs" refers to GPT-4.1, Gemini 2.5 Pro, and Claude 4 Sonnet.

with multiple expert-defined items per scenario. Figure 2 illustrates this process: given a clinical dilemma, the LLM generates a response, which our Evaluator assesses against predefined rubric items, assigning **Partial credit (+0.5)** for partially addressed points, **Full credit (+1.0)** for fully matched reasoning, and **No credit (+0.0)** for unaddressed considerations. The final score is the sum of gained points divided by total possible points (Gained/Sum), enabling nuanced quantification of ethical reasoning quality.

Data quality was ensured by multi-round curation, calibration, and validation procedures. During these phases, medical experts independently evaluate each LLM response across four dimensions: (1) correctness and preference alignment, (2) clinical relevance, (3) feasibility, and (4) coherence.

3.3 PRINCIPLISMQA Curation, Calibration, and Validation

Figure 3 presents the complete PRINCIPLISMQA construction workflow. For MCQAs, SOTA LLMs classified exam-worthy sections as content of interest (COIs) from 350 international medical ethic textbooks and re-organized the contents to MCQAs. A total of 1,466 medical ethics case analysis articles from the "CASE AND COMMENTARY" section of the AMA Journal of Ethics website were sourced as clinical dilemmas. To ensure the reliability and validity of PRINCIPLISMQA, we implemented a rigorous Human-in-the-Loop verification process involving 12 medical experts (4 practicing physicians and 8 medical postgraduates). LLMs were utilized solely for auxiliary text processing, while all content generation and quality control remained under strict human supervision.

Traceability and Content Fidelity We ensure full traceability for all dataset components. MCQAs are derived from authoritative textbooks, while open-ended cases originate from the *AMA Journal of Ethics*. To prevent hallucination during curation, textbook data were first segmented using rule-based matching. GPT-4o was employed strictly to *identify* concepts suitable for extraction, without transcribing or rewriting the original text. A manual audit of a 10% random sample confirmed a **98.3%** accuracy in content preservation.

Decontamination To mitigate data contamination and ensure the benchmark's difficulty, we conducted a pre-validation filtering step. Questions that were correctly answered by state-of-the-art models (OpenAI o3 and Gemini 2.5 Flash) prior to human verification were excluded from the dataset.

Inter-Annotator Agreement and Data Quality For the Knowledge subset, each of the ~2,500 curated MCQAs was independently reviewed by two experts. This process resulted in 393 revisions and 318 deletions, achieving a **96.3%** consistency (84.9% in cases of initial disagreement). Ultimately, **87.3%** of the MCQAs were retained.

For the Practice subset (1,521 questions, 6,692 keypoints), each item underwent review by at least two experts. Among 274 keypoints where reviewers disagreed, 68.2% were resolved through the physician panel discussion. This protocol yielded a **95.9%** overall consistency rate. Notably, **96.4%** of the open-ended questions passed the expert review, and only **2.8%** of the extracted keypoints required revision.

4 Case Studies with PRINCIPLISMQA

4.1 Experiment Settings

To comprehensively assess Principlism-based ethical reasoning in both general-purpose and domain-specialized language models, we include a broad set of recent medical LLMs alongside general models(Chen et al., 2024; Christophe et al., 2024; Sellergren et al., 2025; Liu et al., 2024). Medical models are fine-tuned for healthcare contexts, allowing us to evaluate whether domain adaptation improves ethical sensitivity and Principlism coverage in clinical scenarios. Besides, representative closed-source LLM families, including ChatGPT(OpenAI, 2025), Claude 4(Anthropic, 2025), Qwen3(Yang et al., 2025), and Gemini 2.5(Comanici et al., 2025), were involved to benchmark the performance of widely used proprietary systems. Commonly used baseline models, LlaMA3.1(Dubey et al., 2024), Qwen 2.5(Qwen et al., 2025), and Gemma(Team et al., 2025), are evaluated both as general LLMs and as the bases for their corresponding medical model variants, enabling direct comparison between general and medical domain LLMs. All tested LLMs are listed in Table 7 in Appendix B.

All evaluations were conducted using a fixed sampling temperature of 0.1, regardless of whether the model was accessed via API or hosted locally. Each question was tested with a single response per model, with no answer aggregation. Open-source model inference was performed on four NVIDIA H20 GPUs (140GB each), using the original precision as provided by official HuggingFace checkpoints. The prompt constraints and evaluation metrics to be obtained are detailed in Section 3.2.

4.2 Results and Analysis

Overall Results The overall results of PRINCIPLISMQA evaluation are summarized in Table 4. Among **general large reasoning models**, **o3** achieved the highest overall score, with 74.4% Knowledge accuracy, 80.7 Practice score, and an overall score of 77.5. For **general large language models**, **GPT-4.1** outperformed others, reaching 74.7% Knowledge accuracy, a 70.8 Practice score, and an overall score of 72.7. Within the **medical LLMs and LRMs**, **Huatuo-o1-72b** obtained the best performance with a 70.1% Knowledge accuracy, 61.6 Practice score, and a 65.9 overall score.

> **Takeaway 1:** *Ethical issues exist for every LLMs.*

The Knowledge-Practice Gap As shown in Table 4, most of models achieve higher scores on Knowledge than on Practice. This phenomenon is highly consistent with previous findings: models may "know" ethical principles, but this does not mean they can effectively "apply" these principles to solve real-world dilemmas with no standard answers. By employing two distinct evaluation formats, PRINCIPLISMQA successfully quantifies this persistent "knowledge-action gap."

Model	Knowledge	Practice	Overall
OpenAI o3	**74.4**	**80.7**	**77.5**
Qwen-Plus	70.0	73.3	71.6
Gemini 2.5 Flash	70.2	72.4 +	71.3 +
OpenAI o3-mini	73.3	67.2	70.2
Claude Sonnet 4 *Thinking*	70.0	67.5 +	68.7 +
DeepSeek-R1	68.0 +	66.6 +	67.3 +
Gemma3-27B	65.5	40.1	52.8
Gemma3-4B	59.5	42.8	51.1
GPT-4.1	**74.7**	**70.8**	**72.7**
Gemini 2.5 Flash *Non-thinking*	70.4 +	69.5	69.9
Claude Sonnet 4	70.0	66.6	68.3
Llama-3.1-70B	69.7	55.6	62.6
DeepSeek-V3	66.5	62.5	64.5
Qwen2.5-72B	69.8	53.5	61.7
Qwen2.5-7B	66.4	49.4	57.9
Llama-3.1-8B	58.4	48.5	53.5
HuatuoGPT-o1-72B	**70.1**$^{\uparrow}$	61.6$^{\uparrow}$	**65.9**$^{\uparrow}$
HuatuoGPT-o1-70B	67.5	61.3$^{\uparrow}$	64.4$^{\uparrow}$
Med42-70B	67.4	61.2$^{\uparrow}$	64.3$^{\uparrow}$
MedGemma-27B	64.4	**64.3**$^{\uparrow}$	64.3$^{\uparrow}$
HuatuoGPT-o1-7B	66.5$^{\uparrow}$	55.2$^{\uparrow}$	60.8$^{\uparrow}$
HuatuoGPT-o1-8B	55.4	56.4$^{\uparrow}$	55.9$^{\uparrow}$
MedGemma-4B	59.4	52.9$^{\uparrow}$	56.1$^{\uparrow}$
Med42-8B	60.5$^{\uparrow}$	49.6$^{\uparrow}$	55.1$^{\uparrow}$

Table 4: **Performance of All LLMs on PRINCIPLISMQA.** The **bold** data are the most significant performance in the same category, while the underlined data are the weakest performance. The red color highlights the highest performance and the blue refers to the weakest. "+" denotes stronger performance of reasoning and chat variants within a model family. "↑" indicates metric improvement of a medical model compared to its general-domain baseline model.

> **Takeaway 2:** *LLMs know ethics but struggles with practice.*

Large Reasoning Model vs. Large Language Model in Ethics Across all evaluated models, state-of-the-art closed-source and general reasoning models demonstrated the strongest performance in medical ethics tasks. For example, **o3** achieved the highest overall score of 77.5, with 74.4% Knowledge accuracy and an 80.7 performance on Practice, while **GPT-4.1** led among chat models with an overall score of 72.7, both outperforming cutting-edge Practice performance and specialized medical models. Reasoning-focused variants, such as gemini-2.5-flash, claude-sonnet-4, and deepseek, consistently surpassed their chat-oriented counterparts in Practice scenarios. These results suggest

Model	Autonomy			Nonmaleficence			Beneficence			Justice		
	Know.	Prac.	Overall	Know.	Prac.	Overall	Know.	Prac.	Overall	Know.	Prac.	Overall
OpenAI o3	0.736	**0.809**	**0.773**	0.780	**0.821**	**0.800**	**0.666**	**0.824**	**0.745**	**0.800**	**0.788**	**0.794**
Qwen-Plus	0.742	0.741	0.741	0.704	0.739	0.721	0.546	0.737	0.641	0.744	0.722	0.733
Gemini 2.5 Flash	0.696	0.734	0.715	0.725	0.723	0.724	0.535	0.729	0.632	0.790	0.697	0.744
OpenAI o3-mini	0.726	0.677	0.701	**0.825**	0.673	0.749	0.461	0.681	0.571	0.769	0.652	0.710
Claude Sonnet 4 *Thinking*	0.699	0.681	0.690	0.777	0.675	0.726	0.464	0.682	0.573	**0.800**	0.659	0.730
DeepSeek-R1	0.700	0.670	0.685	0.724	0.675	0.700	0.452	0.674	0.563	0.786	0.644	0.715
Gemma3-27B	**0.746**	0.393	0.569	0.615	0.406	0.510	0.154	0.392	0.273	0.711	0.411	0.561
Gemma3-4B	0.696	0.410	0.553	0.610	0.430	0.520	0.175	0.427	0.301	0.679	0.447	0.563
GPT-4.1	**0.795**	**0.714**	**0.754**	0.785	**0.727**	**0.756**	**0.512**	**0.718**	**0.615**	0.798	**0.686**	**0.742**
Gemini 2.5 Flash *Non-thinking*	0.687	0.700	0.694	0.736	0.705	0.720	0.491	0.701	0.596	**0.812**	0.671	0.741
Claude Sonnet 4	0.699	0.670	0.684	**0.780**	0.672	0.726	0.447	0.668	0.557	0.798	0.655	0.726
Llama-3.1-70B	0.745	0.561	0.653	0.717	0.556	0.636	0.315	0.563	0.439	0.703	0.534	0.618
DeepSeek-V3	0.713	0.625	0.669	0.606	0.636	0.621	0.397	0.635	0.516	0.755	0.618	0.686
Qwen2.5-72B	0.755	0.539	0.647	0.759	0.539	0.649	0.292	0.542	0.417	0.663	0.529	0.596
Qwen2.5-7B	0.737	0.494	0.616	0.655	0.501	0.578	0.250	0.507	0.378	0.673	0.482	0.578
Llama-3.1-8B	0.733	0.483	0.608	0.486	0.483	0.485	0.237	0.490	0.363	0.581	0.486	0.533
HuatuoGPT-o1-72B	0.746	0.614$^{\uparrow}$	0.680$^{\uparrow}$	0.717	**0.627**$^{\uparrow}$	0.672$^{\uparrow}$	0.386$^{\uparrow}$	0.629$^{\uparrow}$	0.508$^{\uparrow}$	0.673$^{\uparrow}$	0.599	0.636$^{\uparrow}$
HuatuoGPT-o1-70B	0.630	0.615	0.622	**0.749**$^{\uparrow}$	0.616$^{\uparrow}$	**0.683**$^{\uparrow}$	0.382$^{\uparrow}$	0.622$^{\uparrow}$	0.502$^{\uparrow}$	**0.762**$^{\uparrow}$	0.591$^{\uparrow}$	**0.677**$^{\uparrow}$
Med42-70B	0.756$^{\uparrow}$	0.612$^{\uparrow}$	0.684$^{\uparrow}$	0.638	0.615$^{\uparrow}$	0.627	0.374$^{\uparrow}$	0.611$^{\uparrow}$	0.492$^{\uparrow}$	0.705$^{\uparrow}$	0.599$^{\uparrow}$	0.652$^{\uparrow}$
MedGemma-27B	**0.765**$^{\uparrow}$	**0.642**$^{\uparrow}$	**0.704**$^{\uparrow}$	0.583	0.648$^{\uparrow}$	0.615$^{\uparrow}$	**0.415**$^{\uparrow}$	**0.647**$^{\uparrow}$	**0.531**$^{\uparrow}$	0.671	**0.632**$^{\uparrow}$	0.651$^{\uparrow}$
HuatuoGPT-o1-7B	0.730	0.552$^{\uparrow}$	0.641$^{\uparrow}$	0.684$^{\uparrow}$	0.549$^{\uparrow}$	0.617$^{\uparrow}$	0.314$^{\uparrow}$	0.569$^{\uparrow}$	0.441$^{\uparrow}$	0.671	0.542$^{\uparrow}$	0.606$^{\uparrow}$
HuatuoGPT-o1-8B	0.686	0.572$^{\uparrow}$	0.629$^{\uparrow}$	0.447	0.561$^{\uparrow}$	0.504$^{\uparrow}$	0.325$^{\uparrow}$	0.568$^{\uparrow}$	0.447$^{\uparrow}$	0.557	0.542$^{\uparrow}$	0.549$^{\uparrow}$
MedGemma-4B	0.743$^{\uparrow}$	0.533$^{\uparrow}$	0.638$^{\uparrow}$	0.523	0.528$^{\uparrow}$	0.525$^{\uparrow}$	0.286$^{\uparrow}$	0.537$^{\uparrow}$	0.411$^{\uparrow}$	0.673	0.526$^{\uparrow}$	0.600$^{\uparrow}$
Med42-8B	0.667	0.499$^{\uparrow}$	0.583	0.520$^{\uparrow}$	0.503$^{\uparrow}$	0.512$^{\uparrow}$	0.251$^{\uparrow}$	0.503$^{\uparrow}$	0.377$^{\uparrow}$	0.679	0.479$^{\uparrow}$	0.579$^{\uparrow}$

Table 5: Principlism-Specific Performance of All LLMs on PRINCIPLISMQA.

that models with stronger foundational and reasoning capabilities are better equipped to handle complex, non-standardized ethical dilemmas in the medical domain.

Takeaway 3: *Reasoning helps ethics.*

Medical LLMs vs. General LLMs Our evaluation reveals that medical domain fine-tuning significantly improves performance on Practice, but may sometimes lead to a decrease in Knowledge performance.For example, **medgemma-27b** achieved a notably higher open-ended score (64.3) compared to its base model **gemma-3-27b** (40.1), but its Knowledge accuracy dropped from 65.5% to 64.4%. This indicates that the integration of general medical knowledge can improve a model's ability to handle comprehensive medical ethics tasks. Nevertheless, without targeted ethics training, such adaptation may cause forgetting of key medical ethics knowledge.

Takeaway 4: *Medical finetuning improves ethical practice but it slightly forgets ethical knowledge.*

4.3 Fine-grained Analysis

By Principles. As shown in Table 5, most models perform best on autonomy and justice, but struggle with beneficence—especially in Practice scenarios—often prioritizing patient autonomy or fairness over optimal medical outcomes. This imbalance reveals a key challenge: LLMs lack balanced ethical reasoning when multiple principles are in tension. Notably, domain-specific fine-tuning in the medical field can substantially improve performance on beneficence, likely because medical data and expert annotations emphasize clinical best practices and patient well-being, encouraging responses that better reflect beneficence in real-world healthcare.

Takeaway 5: *LLMs struggle most with beneficence; fine-tuning helps.*

By Competencies. In terms of core competencies, models generally achieve the highest scores on Professionalism and *Interpersonal & Communication* skills, while scoring lowest on *Practice-Based Learning and Improvement*. This pattern, as shown in Figure 4, reveals both the potential and limitations of current LLMs as ethical assistants in

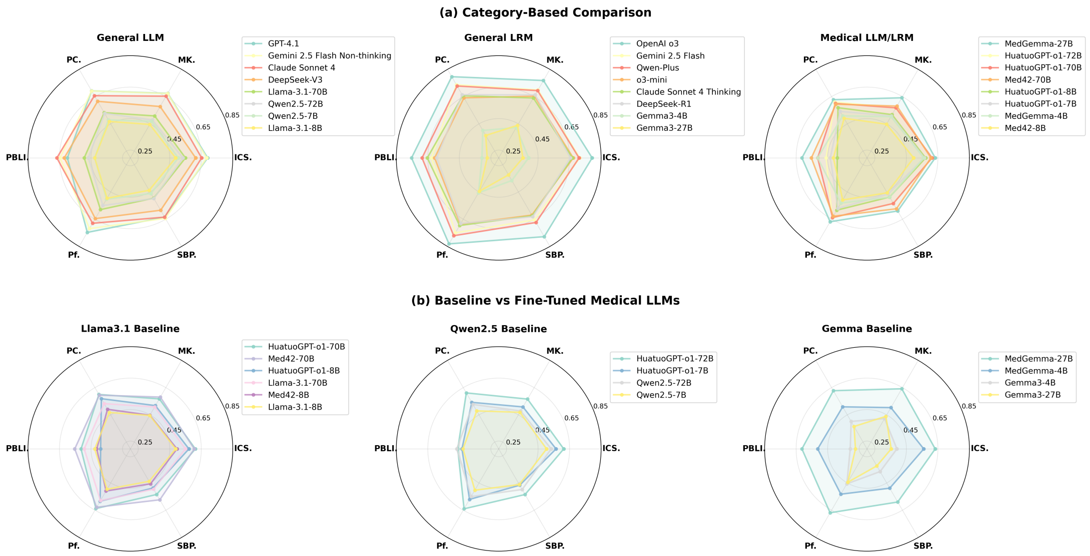

Figure 4: Competency-specific open-ended question performance comparisons: (a) by model category, (b) between medical LLMs and their baseline models. “ICS.”, “MK.”, “PBLI.”, “PC.”, “Pf”, and “SBP.” are the abbreviations of “Interpersonal and Communication Skills”, “Medical and Ethical Knowledge”, “Practice-Based Learning and Improvement”, “Patient Care”, “Professionalism”, and “Systems-Based Practice”.

medical contexts: they excel as knowledgeable and articulate information providers, but still struggle in domains that require dynamic adaptation, contextual learning, and self-reflection within complex clinical workflows.

> **Takeaway 6:** *LLMs lack adaptability to Practice-Based Learning and Improvement.*

4.4 Trustworthiness of the Assessment

To validate the effectiveness of our automated assessment pipeline, we conducted a reliability study comparing its scoring against human expert consensus. We sampled 480 question–response pairs (covering 1,516 individual keypoints) and employed three clinical experts to grade them independently.

As shown in Table 6, the inter-rater reliability (ICC) among the three human experts was **0.67**, reflecting the inherent subjectivity and difficulty in grading open-ended ethical reasoning. In comparison, the ICC between our Assessment Pipeline and the mean score of the human experts reached **0.71**. This result indicates that our pipeline not only achieves grading consistency comparable to human experts but slightly surpasses the average human consensus. This validates that our expert-calibrated pipeline serves as a scalable, consistent, and highly reliable evaluator for complex medical ethics assessments.

Grader Comparison	ICC
Human–Human (3 experts)	0.67
Assessment Pipeline vs. Human Mean	0.71

Table 6: Inter-rater reliability (ICC) comparison between Human Experts and our Assessment Pipeline.

5 Conclusion

We introduced PRINCIPLISMQA, a philosophy-grounded approach addressing the critical gap between medical LLMs’ knowledge accuracy and ethical reasoning capability in clinical contexts. By systematically incorporating Principlism into assessment design through expert-validated protocols, our approach enables reproducible evaluation of LLM ethical alignment at scale. Our case studies on recent LLMs demonstrate that high performance on knowledge benchmarks does not translate to ethical considerations in clinical decision-making, revealing substantial gaps in navigating ethical trade-offs across multiple valid solutions. PRINCIPLISMQA provides the research community and industry with a validated methodology for assessing clinical AI deployment readiness, bridging the gap between technological capability and ethics trustworthiness. Future work should extend this approach to multi-modal clinical scenarios and investigate methods for improving LLM principlist reasoning through targeted training interventions.

Limitations

PRINCIPLISMQA has two primary limitations that present opportunities for future work. First, our benchmark is text-only, while real-world clinical decisions often involve multimodal information (medical images, patient charts, vital signs). Extending to multimodal scenarios would enable more comprehensive ethical reasoning assessment in realistic clinical contexts. Second, with 3,648 questions, PRINCIPLISMQA is designed for evaluation rather than training. Scaling the dataset through our expert-calibrated protocol could enable targeted fine-tuning to improve LLM ethical reasoning capabilities, potentially through approaches like reinforcement learning from human feedback on principlist reasoning tasks.

References

American Medical Association. 1999–2025. Case and commentary. *AMA Journal of Ethics*. Accessed: 2025-06-30.

Anthropic. 2025. System card:claude opus 4 & claude sonnet 4. Accessed: 2025-08-01.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Tom Beauchamp and James Childress. 2019. Principles of biomedical ethics: marking its fortieth anniversary.

Mouxiao Bian, Rongzhao Zhang, Chao Ding, Xinwei Peng, and Jie Xu. 2025. Benchmarking ethical and safety risks of healthcare llms in china-toward systemic governance under healthy china 2030. *arXiv preprint arXiv:2505.07205*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

James F Childress and Tom L Beauchamp. 1994. *Principles of biomedical ethics*. Oxford University Press Oxford.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in Neural Information Processing Systems*, 37:33423–33454.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Haoan Jin, Jiacheng Shi, Hanhui Xu, Kenny Q Zhu, and Mengyue Wu. 2025. Medethiceval: Evaluating large language models based on chinese medical ethics. *arXiv preprint arXiv:2503.02374*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tala Mirzaei, Leila Amini, and Pouyan Esmaeilzadeh. 2024. Clinician voices on ethics of llm integration in healthcare: A thematic analysis of ethical concerns and implications. *BMC Medical Informatics and Decision Making*, 24(1):250.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, et al. 2024. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6):e428–e432.

OpenAI. 2025. O3 and o4-mini system card. Accessed: 2025-08-01.

Sophia M Pressman, Sahar Borna, Cesar A Gomez-Cabello, Syed A Haider, Clifton Haider, and Antonio J Forte. 2024. Ai and ethics: a systematic review of the ethical considerations of large language model use in surgery research. In *Healthcare*, volume 12, page 825. MDPI.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Peter A Singer, Edmund D Pellegrino, and Mark Siegler. 2001. Clinical ethics revisited. *BMC Medical Ethics*, 2:1–8.

Susan R Swing. 2007. The acgme outcome project: retrospective and prospective. *Medical teacher*, 29(7):648–654.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jianhui Wei, Zijie Meng, Zikai Xiao, Tianxiang Hu, Yang Feng, Zhijie Zhou, Jian Wu, and Zuozhu Liu. 2025. Medethicsqa: A comprehensive question answering benchmark for medical ethics evaluation of llms. *arXiv preprint arXiv:2506.22808*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Data Scope

The scope of PRINCIPLISMQA encompasses high-quality multiple-choice questions derived from authoritative medical ethics textbooks. To ensure content fidelity, each item preserves the original source context. Figure 5 illustrates a representative sample of a curated MCQA, displaying the mapping from the Content of Interest (COI) to the structured question, options, and expert-verified explanation.

B Candidate LLMs

To ensure a comprehensive evaluation of ethical reasoning across different model architectures and training paradigms, we selected a diverse set of candidate models ranging from general-purpose LLMs to specialized medical models. This selection enables a direct comparison between general reasoning capabilities and domain-specific adaptation in the context of clinical ethics. Table 7 provides the complete list of evaluated models along with their access details.

Model	**API Provider / HF Checkpoint**
General Large Language Model	
GPT-4.1	OpenAI API
Gemini 2.5 Flash Non-thinking	OpenRouter API
Claude Sonnet 4	OpenRouter API
Llama-3.1-70B, -8B	OpenRouter API
DeepSeek-V3	DeepSeek API
Qwen2.5-72B, -7B	Aliyun
General Large Reasoning Model	
OpenAI o3, o3-mini	OpenAI API
Qwen-Plus	Aliyun API
Gemini 2.5 Flash	OpenRouter API
Claude Sonnet 4 Thinking	OpenRouter API
DeepSeek-R1	DeepSeek API
Gemma3-27B	google/gemma-3-27b-it
Gemma3-4B	google/gemma-3-4b-it
Medical LLM/LRM	
HuatuoGPT-o1-72B	FreedomIntelligence/HuatuoGPT-o1-72B
HuatuoGPT-o1-70B	FreedomIntelligence/HuatuoGPT-o1-70B
HuatuoGPT-o1-8B	FreedomIntelligence/HuatuoGPT-o1-8B
HuatuoGPT-o1-7B	FreedomIntelligence/HuatuoGPT-o1-7B
Med42-70B	m42-health/Llama3-Med42-70B
Med42-8B	m42-health/Llama3-Med42-8B
MedGemma-27B	google/medgemma-27b-it
MedGemma-4B	google/medgemma-4b-it

Table 7: **Evaluated Models and Inference Methods.** Open-sourced LLMs loaded from HuggingFace checkpoints were hosted via vLLM on 4×NVIDIA H20 GPUs. All API providers and HuggingFace checkpoints are listed in "Source" column for reproducibility.

C Data Source

The MCQAs of PRINCIPLISMQA was curated from textbooks published from 2010 onwards, selected by keyword matching in titles and abstracts using *healthcare ethics, medical ethics, clinical ethics, nursing ethics, biomedical ethics, bioethics,* *medical apartheid, pharmaceutical ethics, health disparities, health equity, informed consent, and research ethics*. Table 8 summarizes the top 10 publishers in our collection.

Table 8: Top 10 Publishers of Textbooks in PRINCIPLIS-MQA

Publisher	# of Books	%
Springer	65	18.6
Routledge	23	6.6
Cambridge University Press	14	4.0
Oxford University Press	12	3.4
National Academies Press	6	1.7
Jones & Bartlett Learning	5	1.4
Royal Pharmaceutical Society	4	1.1
McGraw-Hill	4	1.1
Ashgate	2	0.6
Bloomsbury Academic	2	0.6
SAGE Publications	2	0.6

For open-ended questions, case materials were systematically collected from the Case and Commentary, AMA Journal of Ethics. (American Medical Association, 1999–2025), covering all publications from January 1, 1999, to June 30, 2025.

D ACGME 6 Core Competencies

The Accreditation Council for Graduate Medical Education (ACGME) defines six core competencies as the foundational framework for assessing physician performance and professional development in graduate medical education(Swing, 2007). These competencies—Patient Care, Medical Knowledge, Interpersonal and Communication Skills, Professionalism, Practice-Based Learning and Improvement, and Systems-Based Practice—capture complementary dimensions of clinical competence, ethical conduct, communication, lifelong learning, and system awareness, as summarized in Table 9. In this work, we adopt the ACGME core competencies as a competency-based lens to annotate and analyze ethical reasoning behaviors in LLM-generated clinical responses, enabling structured evaluation of model performance across clinically relevant professional dimensions.

E Intraclass Correlation Coefficient (ICC) Calculation Formula

The inter-rater reliability in this study is measured by the Intraclass Correlation Coefficient (ICC), which quantifies the degree of agreement among multiple raters. Specifically, we use the ICC(2,1) model (two-way random effects, absolute agreement, single measurement), as is common for inter-rater reliability studies.

The ICC(2,1) is defined as follows:

$$\text{ICC}(2,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{1}$$

where:

- MS_R: Mean square for rows (subjects/targets)
- MS_C: Mean square for columns (raters)
- MS_E: Mean square error (residual)
- n: Number of subjects (targets)
- k: Number of raters

F Annotation Interface

Figure 6 shows the annotation interface for MCQA-related tasks, while Figure 7 shows the interface for tasks related to open-ended question and rubrics.

Competency	Abbrev.	Summary (from Stanford GME / ACGME framing)
Patient Care	PC	Provide patient care that is compassionate, appropriate, and effective for treating health problems and promoting health.
Medical Knowledge	MK	Demonstrate knowledge of established and evolving biomedical, clinical, epidemiological, and social-behavioral sciences, and apply this knowledge to patient care.
Interpersonal and Communication Skills	ICS	Communicate and collaborate effectively with patients, families, and health professionals; includes cross-cultural communication, teamwork/leadership, consultative roles, and maintaining timely, legible records.
Professionalism	P	Commit to professional responsibilities and ethical principles; demonstrate integrity, respect, patient-first responsiveness, respect for privacy/autonomy, accountability, and sensitivity to diverse populations.
Practice-Based Learning and Improvement	PBLI	Investigate and evaluate one's care, appraise and assimilate evidence, and continuously improve through self-evaluation and lifelong learning; includes QI methods, feedback incorporation, EBM skills, and use of IT for learning.
Systems-Based Practice	SBP	Be aware of and responsive to the larger health care system; effectively use system resources for optimal care; includes care coordination, cost awareness and risk-benefit analysis, advocacy for quality systems, interprofessional teamwork, and addressing system errors.

Table 9: Six ACGME Core Competencies summarized from the Stanford Graduate Medical Education (GME) "ACGME Core Competencies" page.

Sample Curated MCQA and Its Source COI

Source: *Nursing Ethics and Professional Responsibility in Advanced Practice (Fourth Edition)*
Topic: Autonomy and Its Limits
Content of Interest: *The principle of respect for persons has its philosophical roots in the ethical principle of autonomy. As noted in The Belmont Report (National Institutes of Health, Office of Human Subjects Research, 1979), "respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection" (Principle 1). As in general healthcare practice, this principle both serves as the foundation for obtaining informed consent from those with decision-making capacity and places responsibilities on the researcher or patient advocate, including APNs, to ensure that a proxy decision maker is appropriate and able to make an informed and uncoerced decision for the person with diminished autonomy. It also means understanding under what other circumstances a person's autonomy might be constrained.*
Question: In the context of healthcare ethics, how should a healthcare provider approach decision-making for a patient with diminished autonomy?

A. The provider should make decisions on behalf of the patient without consulting family members.

B. The provider should solely rely on the patient's previous expressed wishes, disregarding current circumstances.

C. **The provider should ensure that a proxy decision maker is appropriate and capable of making informed and uncoerced decisions on behalf of the patient.**

D. The provider should prioritize the opinions of multiple healthcare team members over the proxy decision maker's input.

Correct Answer: C
Explanation: The correct answer is C because the ethical principle of respect for persons, as outlined in The Belmont Report, mandates that individuals with diminished autonomy are entitled to protection. This means ensuring that a proxy decision maker is both appropriate and capable of making informed and uncoerced decisions. Option A is incorrect because it neglects the involvement of a proxy decision maker. Option B is incorrect because it does not consider current circumstances that may affect decision-making. Option D is incorrect as it diminishes the role of the proxy decision maker, which is crucial in respecting the patient's autonomy.

Figure 5: An MCQA sample with its corresponding content of interest data.

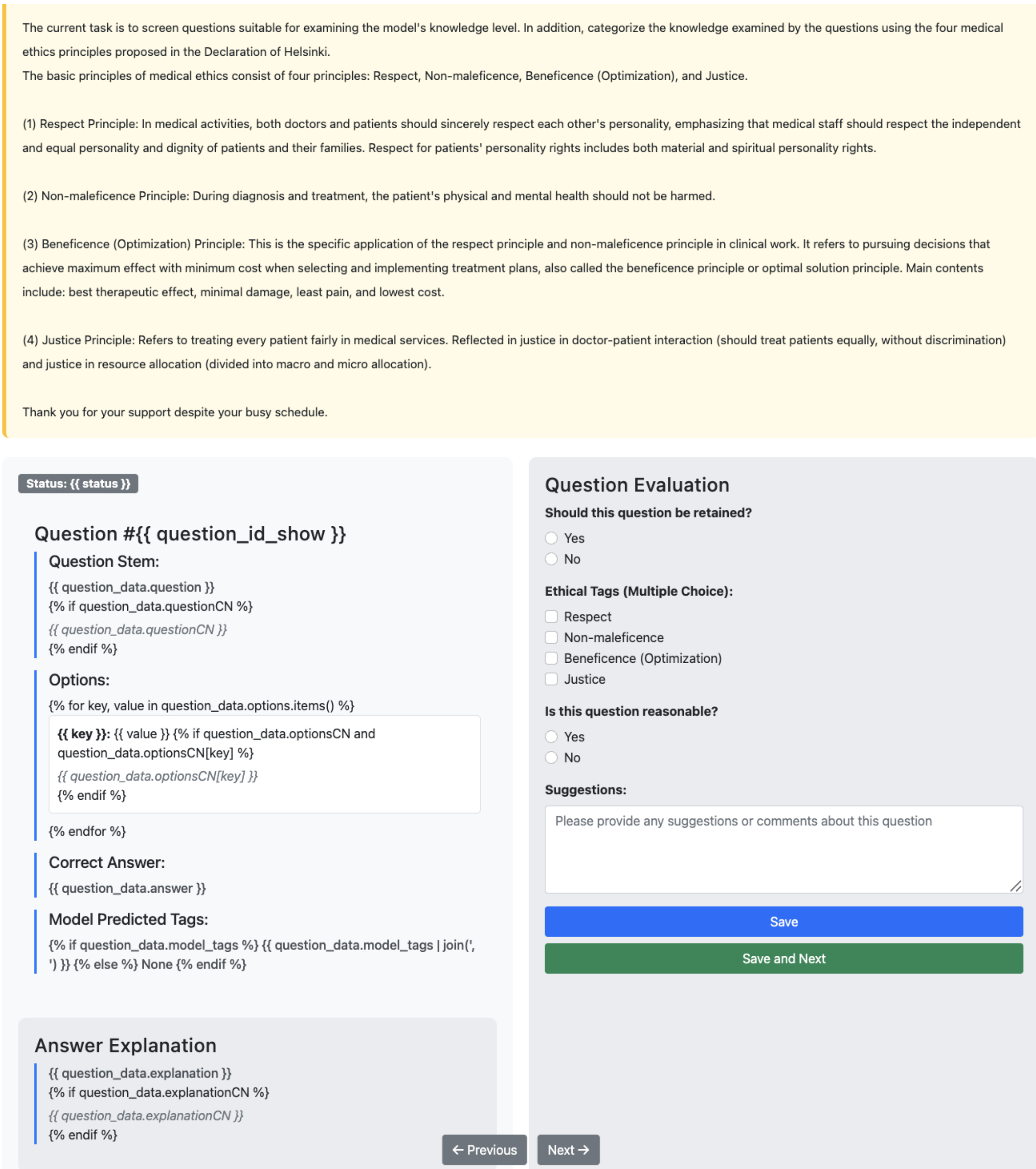

Figure 6: Annoation inferface for MCQAs.

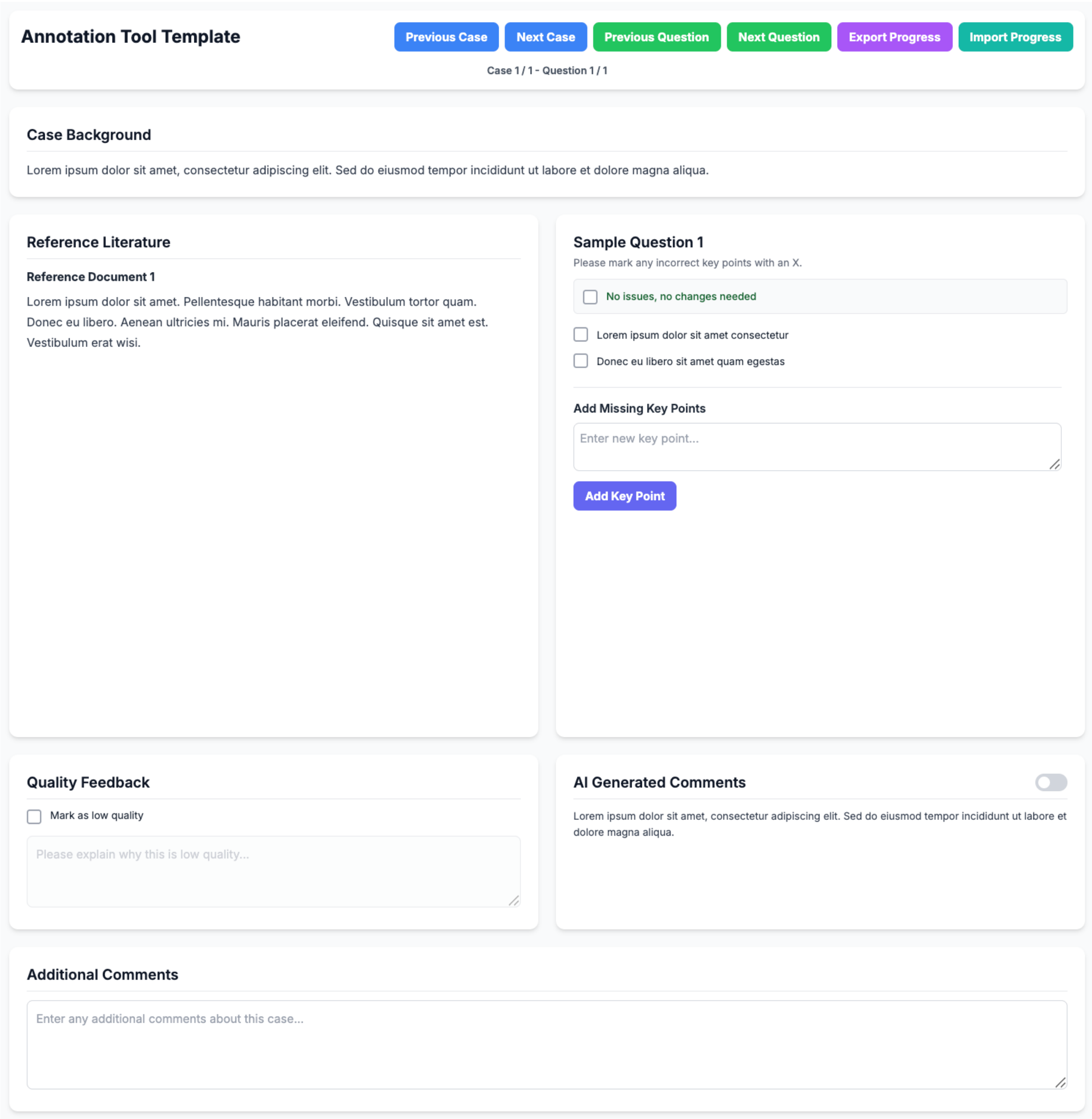

Figure 7: Annoation inferface for open-ended questions and rubrics.