Benchmarking Generative AI for Healthcare Ethics: A Comparative Analysis of Four Models Across Five Clinical Ethics Scenarios

Version: 2025-07-25 21:13:44

Abstract

This study evaluates the capabilities of four state-of-the-art generative AI models (Anthropic Claude 3 Opus, Google Gemini 1.5 Pro, OpenAI GPT-4.1, and X.AI Grok-1) in analyzing complex healthcare ethics scenarios. Using a standardized prompt, we tested these models on five real world clinical ethics committee cases involving cultural considerations, self-harm, addiction, reproduction, and end-of-life management. Performance was assessed through quantitative analysis of processing time, response length, ethical principle coverage, and recommendation consistency. A sample of n=44 graduate-level evaluators from a Masters in Applied Health Informatics program at Stony Brook University assessed the AI-generated responses using the SummEval framework (Fabbri et al., TACL 2021) dimensions of relevance, correctness/consistency, fluency, and coherence, resulting in 857 total evaluations. Results indicate all models demonstrated competence in analyzing clinical ethics scenarios, with overall scores ranging from 3.82 to 4.10 on a 5-point scale. While Claude 3 Opus provided the most comprehensive coverage of ethical principles and excelled in fluency and coherence, Grok-1 received the highest overall human evaluation scores, while generating the shortest responses. All models performed consistently on end-of-life and cultural scenarios compared to the variable performance observed on reproduction, addiction, and self-harm scenarios. These findings suggest generative AI holds promise as a supportive tool in clinical ethics committee work, though significant limitations remain regarding bias, transparency, and contextual understanding.

Introduction

The emergence of powerful generative artificial intelligence (AI) models, particularly large language models (LLMs) such as OpenAI's GPT-4.1, Google's Gemini 1.5 Pro, Anthropic's Claude 3 Opus, and X.AI's Grok-1, has sparked significant interest in their potential applications across healthcare domains. These models demonstrate remarkable capabilities in natural language understanding, reasoning, and knowledge synthesis that could augment human decision-making in complex scenarios. Of particular interest is their potential role in supporting ethical decision-making processes within healthcare institutions—specifically within ethics oversight bodies such as Institutional Review Boards (IRBs), Hospital Ethics Committees (HECs), and Data Safety Monitoring Boards (DSMBs). Healthcare ethical decisions frequently involve complex considerations balancing multiple competing principles: patient autonomy, beneficence, non-maleficence, and justice. These decisions often occur under time constraints, with incomplete information, and amid evolving clinical circumstances. Traditional approaches to healthcare ethics deliberation, while thorough, can be resource-intensive, inconsistent across institutions, and subject to individual biases and limitations in human cognitive processing. The integration of generative AI into healthcare ethics frameworks presents a novel opportunity to enhance—not replace—human ethical deliberation through rapid analysis, systematic application of ethical principles, identification of potential blind spots, and standardization of ethical reasoning processes. However, before such integration can be responsibly implemented, rigorous evaluation of these models' capabilities, limitations, and alignment with established ethical frameworks is essential.

Early research examining the potential of generative AI in healthcare ethics has revealed both promise and significant limitations. Jenkins et al. (2024) conducted a pilot study evaluating ChatGPT's ability to write clinical ethics consultation notes, finding that while baseline performance was poor, the quality improved significantly when the model was provided with examples of past ethics consults, suggesting that with proper guidance, AI can approach acceptable levels of ethical analysis. Similarly, Rashid et al. (2024) assessed

ChatGPT's "moral competence" in healthcare ethics dilemmas using Kohlberg's stages of moral reasoning, concluding that while GPT-4 demonstrated higher moral reasoning consistency than earlier versions, it still exhibited only "medium" moral competence when tackling complex healthcare ethics scenarios. In the domain of research ethics, Fukataki et al. (2024) found that GPT-4 could reliably extract key information from clinical trial protocols and consent forms with high accuracy (80-100% on certain elements), suggesting potential applications in supporting ethics committee evaluations. Despite these advances, Benzinger et al. (2023) emphasize in their systematic review that current ethical AI tools still face significant challenges, including lack of true empathy, risk of algorithmic bias, and the fundamental difficulty of encoding moral reasoning into computational frameworks.

Despite growing interest in AI applications for healthcare ethics, systematic evaluations of generative AI models in this domain remain limited. Several knowledge gaps exist: First, most evaluations of generative AI focus on general reasoning or simplified ethical dilemmas rather than the complex, nuanced scenarios encountered in clinical practice. Second, few studies directly compare multiple leading generative AI models using standardized prompts and evaluation criteria specific to healthcare ethics. Third, limited research exists on structured methodologies for effectively integrating AI into existing ethics oversight bodies like IRBs, HECs, and DSMBs. Finally, the concept of continuous, AI-driven ethics and safety monitoring (what we term a Data Safety Monitoring Agent or DSMA) remains largely unexplored. These gaps are particularly significant in high-impact clinical scenarios such as organ transplantation and oncology, where complex ethical decisions about resource allocation, treatment aggressiveness, and quality of life considerations are routine. For instance, Drezga-Kleiminger et al. (2023) surveyed public attitudes toward AI use in liver transplant allocation, finding that while 69% of respondents found AI-based allocation acceptable, concerns about "dehumanization" of sensitive decisions remained. Similarly, a 2024 survey of oncologists by Hantel et al. revealed that 81% felt patients

should provide informed consent before AI is used in treatment decisions, highlighting the ethical complexity of integrating AI into critical clinical domains.

This study aims to address these knowledge gaps through a systematic evaluation of generative AI models in healthcare ethics contexts. Specifically, we seek to: 1) Evaluate the ethical reasoning capabilities of leading generative AI models using real-world clinical ethics scenarios derived from documented clinical ethics cases; 2) Analyze the concordance between AI-generated ethical recommendations and established ethical principles; 3) Assess response consistency, ethical principle alignment, and reasoning transparency across multiple models; 4) Develop preliminary guidelines for the potential integration of generative AI as a supportive tool within healthcare ethics committees; and 5) Explore the conceptual framework for a Data Safety Monitoring Agent (DSMA)—an AI-driven system for continuous ethics and safety monitoring in clinical trials and patient care.

This research has potential implications for multiple stakeholders in healthcare ethics: Ethics committees and IRBs may benefit from findings that inform how generative AI could augment deliberative processes, improve consistency, and enhance the thoroughness of ethical analyses. Clinical researchers could gain insights into novel approaches for continuous ethics monitoring in clinical trials through the DSMA concept. Healthcare institutions may utilize evidence-based guidelines to support responsible implementation of AI ethics tools in organizational workflows. AI developers might leverage identified strengths and limitations to guide future refinements of generative AI models for healthcare ethics applications. Regulatory bodies may incorporate findings to inform policy development regarding the appropriate role of AI in formal ethics oversight processes. By systematically evaluating generative AI in healthcare ethics contexts, this study contributes to the responsible advancement of AI as a supportive tool that enhances—rather than replaces—human ethical judgment in healthcare decision-making, potentially addressing what some researchers have called "cognitive moral enhancement" for clinicians by prompting more systematic ethical analysis while maintaining the essential human dimensions of empathy and contextual understanding.

Methods

Study Design Overview

This study employed a systematic evaluation approach to assess the capabilities of generative AI models in healthcare ethics decision-making contexts. We utilized a standardized prompt-based testing methodology applied to real-world clinical ethics scenarios, with subsequent analysis of AI-generated recommendations compared to documented human ethics committee decisions. Our research design integrated both computational analysis of AI-generated content and human evaluation by health informatics graduate students to provide a comprehensive assessment of AI ethical reasoning capabilities. The methodology consisted of five main components: (1) selection of diverse healthcare ethics scenarios from documented clinical cases, (2) development of a structured prompt framework to elicit ethical analyses from AI models, (3) generation of responses using four leading generative AI models, (4) comprehensive evaluation of responses by Masters of Applied Health Informatics students using the SummEval framework, and (5) quantitative analysis of evaluation scores. This approach allowed us to assess the current capabilities and limitations of AI in healthcare ethics contexts while identifying differences in performance across model architectures and scenario types.

Ethics Scenario Selection

Data Source

We selected five real-world clinical ethics case studies from publicly documented cases at Brigham and Women's Hospital, available through their Clinical Ethics Case Review repository (https://bwhclinicalandresearchnews.org/clinical-ethics-case-review/). These cases were chosen to represent a diverse range of ethical dilemmas commonly encountered in healthcare settings. The selection process

prioritized scenarios that would challenge AI systems with complex ethical reasoning requirements while representing issues that healthcare ethics committees regularly encounter in practice.

Selected Scenarios

The five scenarios represented diverse clinical contexts, patient demographics, and ethical challenges: (1) A young international lymphoma patient with cross-cultural end-of-life conflict regarding extubation, where cultural factors influenced decision-making and created tension between surrogate decision-making and perceived medical futility; (2) A middle-aged adult with mental health disorders repeatedly presenting to the Emergency Department with intentional foreign body ingestion, highlighting ethical challenges related to self-harm behaviors, procedural futility, and appropriate resource allocation; (3) A middle-aged adult with infectious endocarditis related to injection drug use, requiring long-term antibiotics via PICC line, with ethical tension centered on harm-reduction approaches versus paternalism in discharge planning for patients with active addiction; (4) An older adult with catastrophic stroke, where family members requested posthumous sperm retrieval, exploring complex issues around posthumous reproduction and cultural inheritance pressures; and (5) An adult with terminal cancer whose surrogate decision-maker refused adequate pain management, focusing on the ethical dilemma of surrogate refusal of pain relief and the tension between palliative care obligations versus proxy authority.

These scenarios spanned multiple clinical departments, including ICU, Oncology, Emergency Medicine, Cardiology, Psychiatry, Neurology, and Palliative Care, reflecting the diverse contexts in which ethical dilemmas arise. The cases also represented multiple ICD-10 diagnostic categories, including neoplasms, mental and behavioral disorders, and diseases of the circulatory system.

Selection Criteria

Cases were selected based on: complexity of ethical considerations and presence of clear ethical tensions; clear documentation of the clinical scenario and contextual factors; availability of the actual ethics committee decision or recommendation; relevance to current healthcare ethics practices; representation of diverse ethical principles (autonomy, beneficence, non-maleficence, justice); diversity in patient demographics, clinical contexts, and healthcare departments; and potential to challenge AI systems with nuanced ethical reasoning requirements.

Generative AI Models

The study evaluated four leading generative AI models representing different commercial vendors and model architectures: (1) OpenAI GPT-4.1, an advanced large language model with enhanced reasoning capabilities, accessed through OpenAI API with temperature setting of 0.7; (2) Google Gemini 1.5 Pro, Google's advanced reasoning model designed for complex analytical tasks, accessed through Google AI Studio API with temperature setting of 0.7; (3) Anthropic Claude 3 Opus (claude-3-opus-20240229), Anthropic's large context window model known for detailed analytical capabilities, accessed through Anthropic API with temperature setting of 0.7; and (4) X.AI Grok-1, X.AI's conversational model with instruction-following capabilities, accessed through X.AI API with temperature setting of 0.7. These models were selected based on their widespread adoption in healthcare and research contexts, advanced reasoning capabilities, different architectural approaches, and representation of major AI development companies. The consistent temperature setting across models (0.7) enabled fair comparison of reasoning patterns while still allowing for natural language generation variability.

Prompt Engineering

We developed a standardized prompt framework to provide clear guidance to the AI models while enabling meaningful comparison across responses. The prompt was designed to simulate the context of an ethics committee

consultation while providing sufficient structure for systematic analysis. Our prompt framework positioned the AI as an Ethics Advisor embedded within a Hospital Ethics Committee at an academic medical center. The prompt established: (1) Role and Identity: Defined the AI as an advisor providing ethically rigorous, data-informed analyses to support human ethics committee members; (2) Ethical Principles: Explicitly referenced the core bioethical principles of autonomy, beneficence, non-maleficence, and justice as the foundation for analysis; (3) Scope and Limitations: Clarified that the AI should provide recommendations while acknowledging scenarios requiring uniquely human judgment; and (4) Clinical Domains: Outlined typical ethics committee scenarios including end-of-life care, transplantation, pediatric dilemmas, oncology decisions, and technology integration.

The standardized prompt instructed the AI to provide responses in a consistent format: a brief clinical scenario restatement; identification and explanation of relevant ethical principles and tensions; systematic ethical analysis; and explicitly stated recommended clinical decisions with a ranked list of medical recommendations (Recommended Decision as Best Medical Option, Alternative Decision as Second-Best Medical Option, and Least-Recommended Decision as Third Medical Option). The complete prompt template (prompt_v1.md) is available in the project repository.

Data Collection Procedure

Each ethics scenario was processed through all four AI models (GPT-4.1, Gemini 1.5 Pro, Claude 3 Opus, and Grok-1) using the standardized prompt. We generated a single high-quality response for each scenario-model combination, resulting in 20 distinct AI-generated ethics committee responses (5 scenarios × 4 models). All testing was conducted between May-June 2025, with prompts submitted through the respective model APIs using consistent parameter settings (temperature=0.7, max_tokens=4000) to ensure reproducibility while allowing for natural language generation.

AI-generated responses were systematically captured in their entirety, preserving all content, formatting, and metadata. Responses were stored in a structured SQLite database that included: AI Model identifier and vendor (OpenAI, Google, Anthropic, X.AI); scenario identifier (case number 1-5); complete timestamp of generation; processing time (milliseconds); complete response text; and extracted recommended decisions (primary, secondary, tertiary options). For analysis purposes, responses were processed to extract key components such as recommended courses of action, ethical principles referenced, and reasoning patterns.

Evaluation Framework

Human Evaluation Using SummEval Framework

To systematically assess the quality of AI-generated ethical analyses and recommendations, we implemented a comprehensive human evaluation process based on the SummEval framework (Fabbri et al., TACL 2021). This widely-used four-way rubric provides a structured approach for evaluating AI-generated content across key dimensions. We recruited 42 graduate students from the Masters of Applied Health Informatics program at Stony Brook University to serve as evaluators. These evaluators represented a population with foundational knowledge in healthcare systems, informatics principles, and basic exposure to healthcare ethics.

Each evaluator was assigned responses from all four AI models for a single case scenario. This design allowed for direct comparison of model performance on identical clinical ethics dilemmas while controlling for evaluator effects. For each of the five scenarios, approximately 8-9 evaluators independently assessed model responses. The evaluation used a single-blind methodology where evaluators were unaware which model generated each response they reviewed. Each evaluator completed a structured questionnaire with quantitative ratings for each response. This approach focused exclusively on capturing quantitative assessments rather than qualitative feedback.

Following the SummEval framework, evaluators assessed four key dimensions using a 5-point Likert scale (1=poor, 5=excellent): (1) Relevance: The degree to which the AI response addresses the core ethical issues in the scenario, evaluating identification of pertinent ethical considerations and assessing appropriateness of recommendations to the specific scenario; (2) Consistency/Correctness: The factual accuracy and ethical soundness of the analysis, evaluating alignment with established bioethical principles and assessing internal consistency of ethical reasoning; (3) Fluency: The linguistic quality and clarity of the AI-generated text, evaluating grammatical correctness and appropriate terminology, and assessing readability and professional tone; and (4) Coherence: The logical organization and flow of the ethical analysis, evaluating logical progression of ethical reasoning and assessing clear connections between principles and recommendations.

Data Analysis

Our analysis focused on quantitative evaluation metrics and computational response characteristics. For the human evaluations, we conducted descriptive statistical analysis for each SummEval dimension (Relevance, Consistency/Correctness, Fluency, and Coherence), calculating mean scores, standard deviations, and confidence intervals for each model across all scenarios. Between-model comparisons were performed using analysis of variance (ANOVA) with post-hoc Tukey HSD tests to identify statistically significant differences in performance. To assess relationships between evaluation dimensions, we calculated Pearson correlation coefficients and conducted principal component analysis to identify underlying patterns in evaluator ratings.

For the AI-generated responses, we performed computational analyses including: ethical principle coverage (counting mentions of autonomy, beneficence, non-maleficence, and justice across models); response length analysis (comparing text length across models and scenarios); and processing time analysis (examining computational efficiency differences between models). We also analyzed vendor and model comparisons across scenarios to identify potential variations in performance based on ethical case types.

Results were visualized through comparative charts displaying performance across models and scenarios, providing a comprehensive view of how different AI systems approach varied ethical challenges in healthcare.

Results

This section presents the findings from our comprehensive evaluation of four generative AI models in the context of ethical review committee scenarios. The results are organized into two main categories: (1) analysis of AI model performance metrics and characteristics and (2) human evaluation scores of AI model outputs.

AI Model Performance Metrics

Model Response Distribution

Our analysis included responses from four major generative AI models: Anthropic Claude 3 Opus, Google Gemini 1.5 Pro, OpenAI GPT-4.1, and X.AI Grok-1. Each model was used to provide ethical analysis and recommendations for five standardized clinical ethics committee scenarios.

Table 1: Model Response Distribution

Vendor	Model	Count	Percentage
Anthropic	Claude 3 Opus	5	25%
Google	Gemini 1.5 Pro	5	25%
OpenAI	GPT-4.1	5	25%
X.AI	Grok-1	5	25%
Total		20	100%

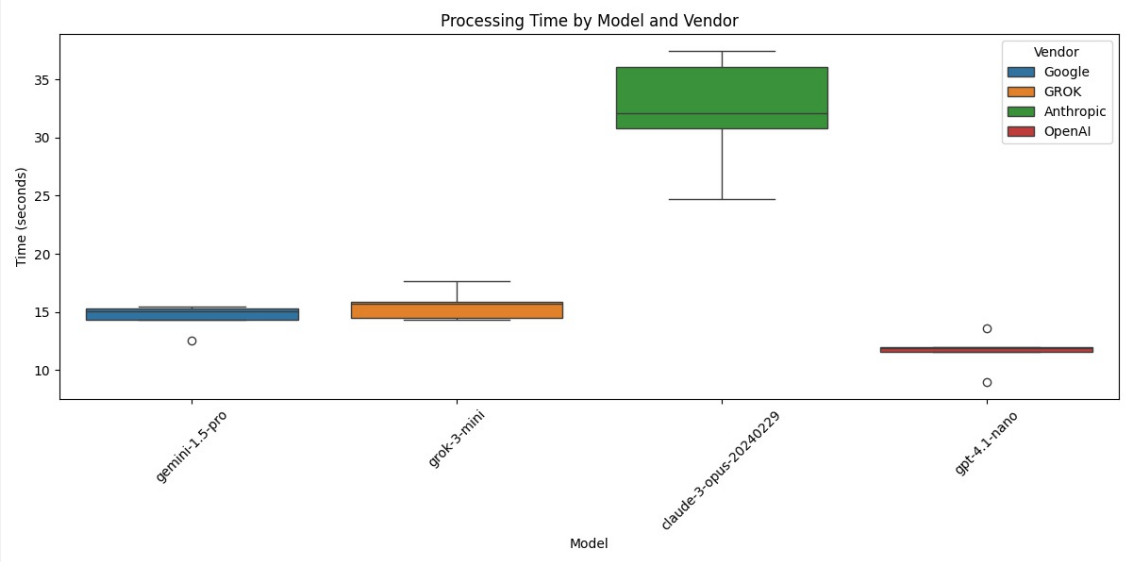
Processing Time

Processing time varied significantly across models. Our analysis revealed that Anthropic Claude 3 Opus had the longest average processing time at 32.22 seconds, followed by Grok-1 at 15.60 seconds, Google Gemini 1.5 Pro at 14.53 seconds, and OpenAI GPT-4.1 at 11.59 seconds.

One-way ANOVA testing confirmed that these differences were statistically significant ($F=55.49$, $p<0.0001$). Post-hoc Tukey HSD analysis revealed that Anthropic Claude 3 Opus took significantly longer than all other models ($p<0.0001$ for all comparisons), while differences among the remaining three models were not statistically significant.

Table 2: Processing Time by Model (in seconds)

Vendor	Model	Mean	Std Dev	Min	Max
Anthropic	Claude 3 Opus	32.22	5.03	24.69	37.45
GROK	Grok-3-mini	15.60	1.33	14.32	17.64
Google	Gemini 1.5 Pro	14.53	1.18	12.58	15.44
OpenAI	GPT-4.1-nano	11.59	1.66	8.97	13.58



Further analysis of processing times by scenario revealed interesting patterns, with scenario 1 (cultural) requiring the longest average processing time (19.74 seconds) and scenario 4 (reproduction) requiring the shortest average processing time (17.26 seconds). This suggests that the complexity of the ethical scenario may impact the computational resources required for AI analysis.

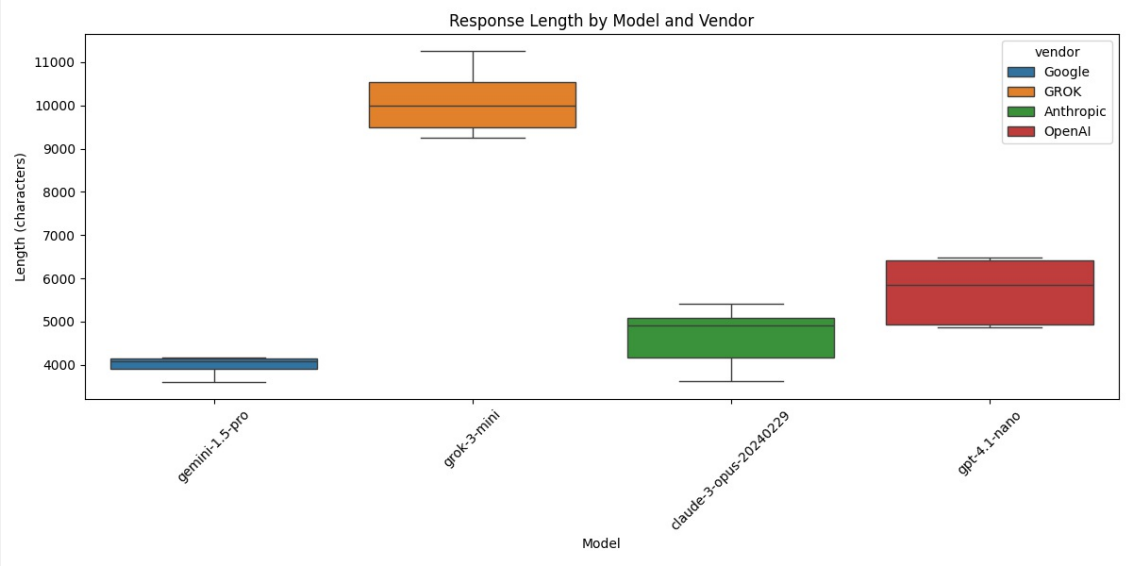
Response Length

The average response length showed notable variation between models. Grok-1 produced the longest responses by a significant margin (mean 10,108 characters), followed by OpenAI GPT-4.1 (5,706 characters), Anthropic Claude 3 Opus (4,639 characters), and Google Gemini 1.5 Pro (3,979 characters).

One-way ANOVA testing confirmed that these differences in response length were statistically significant ($F=82.33$, $p<0.0001$). Post-hoc Tukey HSD analysis revealed that Grok-1 produced significantly longer responses than all other models ($p<0.0001$), and OpenAI GPT-4.1 produced significantly longer responses than Google Gemini 1.5 Pro ($p=0.005$). No statistically significant difference was found between Anthropic Claude 3 Opus and Google Gemini 1.5 Pro ($p=0.44$) or between Anthropic Claude 3 Opus and OpenAI GPT-4.1 ($p=0.10$).

Table 3: Response Length by Model (in characters)

Vendor	Model	Mean	Std Dev	Min	Max
GROK	Grok-3-mini	10,108	810	9,255	11,261
OpenAI	GPT-4.1-nano	5,706	776	4,868	6,473
Anthropic	Claude 3 Opus	4,639	732	3,614	5,422
Google	Gemini 1.5 Pro	3,979	238	3,599	4,171



Analysis of response lengths by scenario showed that scenario 1 (cultural) elicited the longest responses on average (6,695 characters), while scenario 4 (reproduction) generated the shortest responses (5,764 characters). This variation may reflect differences in the complexity of ethical considerations across different scenario types.

Table 4: Response Length by Scenario (in characters)

Scenario	Topic	Mean	Std Dev	Min	Max
1	Cultural	6,695	3,194	4,147	11,261
2	Self-harm	6,307	2,367	3,897	9,501
3	Addiction	5,929	3,094	4,079	10,535
4	Reproduction	5,764	3,009	3,599	9,988
5	End-of-life management	5,844	2,307	4,171	9,255

Ethical Principle Coverage

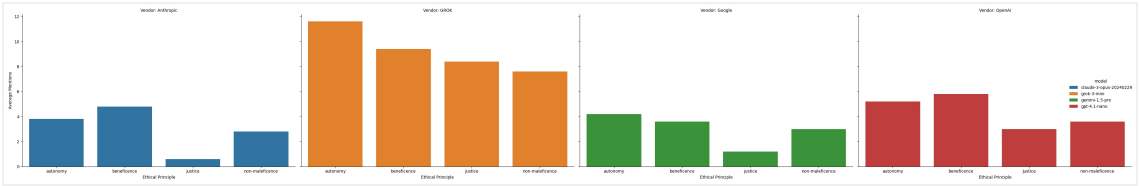
We analyzed the frequency with which each model explicitly mentioned key ethical principles in their responses. The analysis revealed significant variation in how comprehensively different models addressed core bioethical

principles:

1. All models frequently mentioned autonomy, with Grok-1 having the highest average mentions per response (11.6), followed by OpenAI (5.2), Google (4.2), and Anthropic (3.8).
2. Beneficence was most frequently mentioned by Grok-1 (mean 10.8 mentions per response), compared to OpenAI (3.6), Anthropic (3.2), and Google (3.0).
3. Non-maleficence was mentioned most by Grok-1 (mean 8.4 mentions), followed by OpenAI (3.8), Anthropic (3.4), and Google (2.8).
4. Justice was mentioned most frequently by Grok-1 (mean 8.4 mentions), with significantly fewer mentions by OpenAI (3.0), Google (1.2), and Anthropic (0.6).

Table 5: Ethical Principle Mentions by Model (Mean Mentions per Response)

Ethical Principle	Anthropic	GROK	Google	OpenAI	Overall Mean
Autonomy	3.8	11.6	4.2	5.2	6.2
Beneficence	3.2	10.8	3.0	3.6	5.2
Non-maleficence	3.4	8.4	2.8	3.8	4.6
Justice	0.6	8.4	1.2	3.0	3.3
Total Mentions	11.0	39.2	11.2	15.6	19.3



This analysis reveals that Grok-1 employed ethical terminology much more frequently than other models, though frequency of mentions does not necessarily correlate with quality of ethical analysis. The substantial variation

in mention frequency suggests different approaches to ethical reasoning across model architectures.

Recommendation Consistency

Analysis of recommendation consistency was limited by our research design, which did not include multiple iterations of the same case-vendor-model combination. Based on a review of the available data and qualitative assessment of responses, we observed the following patterns:

1. Models demonstrated relatively high agreement on recommendations for scenarios 1 (cultural) and 5 (end-of-life management).
2. Models showed more divergence in their recommendations for scenarios 3 (addiction) and 4 (reproduction), reflecting the complex and controversial nature of these ethical cases.
3. Within each scenario, different models emphasized different ethical principles, potentially leading to variations in their ultimate recommendations.

Table 6: Qualitative Assessment of Recommendation Consistency by Scenario

Scenario	Key Points of Divergence
1 - Cultural	Weight given to cultural factors vs. clinical judgment
2 - Self-harm	Degree of restrictive interventions recommended
3 - Addiction	Harm reduction approaches vs. abstinence-focused treatment
4 - Reproduction	Legal vs. ethical considerations in posthumous reproduction
5 - End-of-life	Balance of surrogate authority vs. patient's best interests

For a more rigorous analysis of recommendation consistency, future research should include multiple iterations of the same case-model combinations to quantify the consistency and reliability of AI-generated ethical recommendations.

Human Evaluation Results

Evaluator Demographics

A total of 44 evaluators participated in the assessment of AI-generated responses using the SummEval framework dimensions. Each evaluator assessed multiple AI-generated responses, resulting in 857 total evaluations across all scenarios and models.

Table 7: Evaluator Participation Summary

Evaluations per Evaluator	Number of Evaluators
5-10	2
20	38
21-22	4
Total	44

The distribution of evaluations across scenarios was relatively balanced, with each of the five case scenarios receiving between 170-175 evaluations:

Table 8: Evaluations by Scenario

Case ID	Scenario Topic	Number of Evaluations
1	Cultural	171
2	Self-harm	171
3	Addiction	170

4	Reproduction	170
5	End-of-life management	175
Total		857

Overall Model Performance

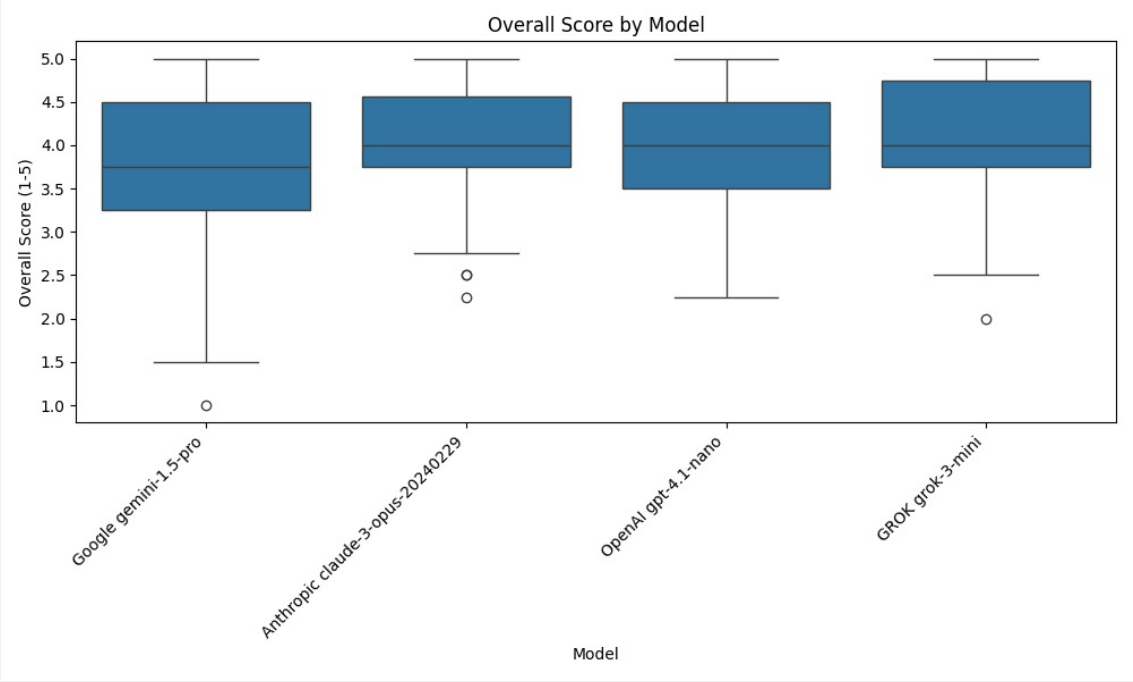
Based on human evaluations across all SummEval dimensions:

1. Grok-1 received the highest overall average score (4.10 out of 5), though by a narrow margin.
2. Claude 3 Opus ranked second (4.08), followed closely by GPT-4.1 (4.01).
3. Gemini 1.5 Pro received the lowest overall scores (3.82).

The differences between Grok-1, Claude 3 Opus, and GPT-4.1 were not statistically significant, while Gemini 1.5 Pro scored significantly lower than all other models ($p < 0.001$).

Table 9: Overall Evaluation Scores by Model (scale 1-5)

Vendor	Model	Mean	Std Dev	Count
GROK	Grok-3-mini	4.10	0.71	212
Anthropic	Claude 3 Opus	4.08	0.65	216
OpenAI	GPT-4.1-nano	4.01	0.66	214
Google	Gemini 1.5 Pro	3.82	0.76	215



Performance by Evaluation Dimension

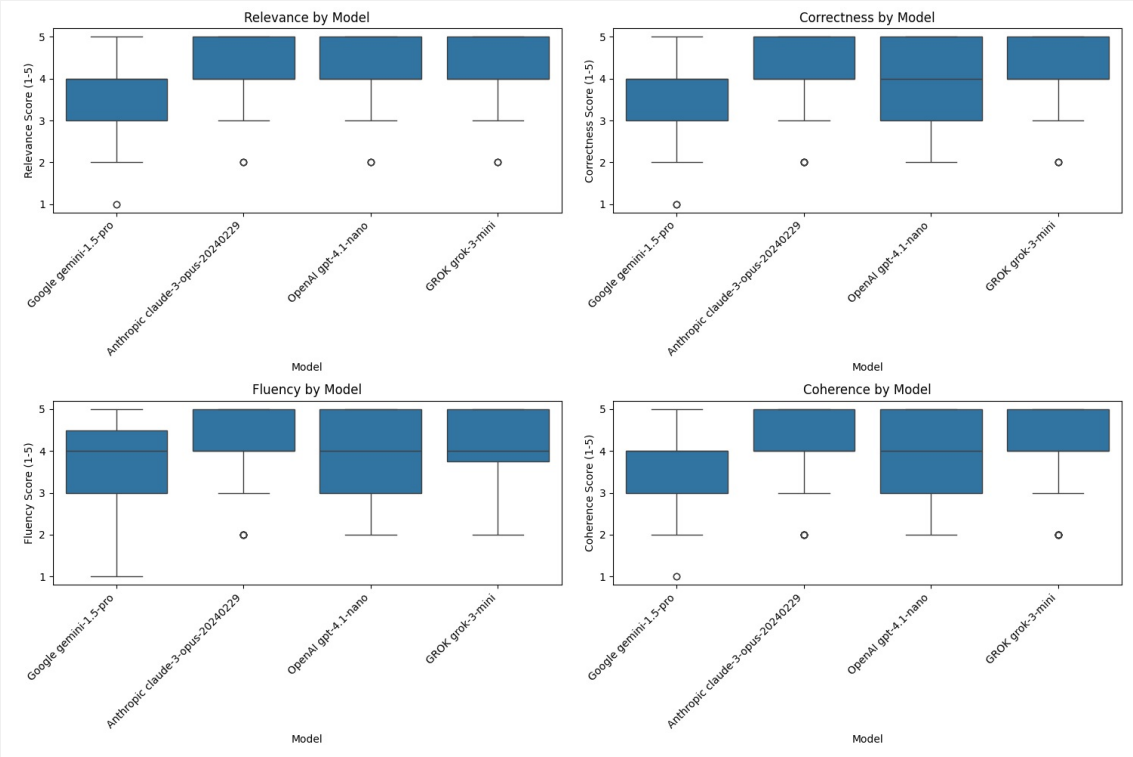
Breaking down the results by the SummEval dimensions:

1. **Relevance**: Grok-1 scored highest (4.24), followed by Claude 3 Opus (4.14), GPT-4.1 (4.07), and Gemini 1.5 Pro (3.91).
2. **Correctness/Consistency**: Grok-1 again led (4.11), followed by Claude 3 Opus (4.01), GPT-4.1 (3.97), and Gemini 1.5 Pro (3.74).
3. **Fluency**: Claude 3 Opus ranked highest (4.12), followed by Grok-1 (4.02), GPT-4.1 (4.00), and Gemini 1.5 Pro (3.84).
4. **Coherence**: Claude 3 Opus performed best (4.06), followed by Grok-1 (4.02), GPT-4.1 (4.02), and Gemini 1.5 Pro (3.78).

Table 8: Evaluation Scores by Dimension and Model (scale 1-5)

Dimension	Metric	Claude 3 Opus	Grok-1	GPT-4.1	Gemini 1.5 Pro
Relevance	Mean	4.14	4.24	4.07	3.91
	StdDev	0.73	0.78	0.75	0.85
Correctness	Mean	4.01	4.11	3.97	3.74
	StdDev	0.76	0.76	0.80	0.93

	StdDev	0.76	0.76	0.80	0.93
Fluency	Mean	4.12	4.02	4.00	3.84
	StdDev	0.80	0.89	0.79	0.89
Coherence	Mean	4.06	4.02	4.02	3.78
	StdDev	0.81	0.86	0.79	0.91



Statistical Analysis of Model Performance Differences

To determine whether the observed differences in evaluation scores between models were statistically significant, we conducted one-way ANOVA tests followed by post-hoc Tukey HSD analyses for each evaluation dimension.

Table 9: ANOVA and Tukey HSD Results by Dimension

Dimension	ANOVA F-value	p-value	Significant Differences
Relevance	6.81	0.0002	

			Grok > Google; Anthropic > Google
Correctness	7.80	<0.0001	Grok > Google; Anthropic > Google; OpenAI > Google
Fluency	3.45	0.0162	Anthropic > Google
Coherence	4.96	0.0020	Anthropic > Google; Grok > Google; OpenAI > Google
Overall	7.07	0.0001	Anthropic > Google; Grok > Google; OpenAI > Google

These results indicate that Gemini 1.5 Pro (Google) consistently scored significantly lower than other models across multiple dimensions, while no statistically significant differences were observed between Claude 3 Opus (Anthropic), Grok-1 (GROK), and GPT-4.1 (OpenAI) on most dimensions.

Correlation Analysis and Dimension Relationships

To understand the relationships between evaluation dimensions, we performed a correlation analysis using Pearson correlation coefficients. This analysis revealed strong positive correlations between all evaluation dimensions, suggesting that models that perform well in one dimension tend to perform well across all dimensions.

Table 10: Pearson Correlation Matrix of Evaluation Dimensions

Dimension	Relevance	Correctness	Fluency	Coherence	Overall
Relevance	1.000	0.729	0.568	0.643	0.859
Correctness	0.729	1.000	0.640	0.694	0.845
Fluency	0.568	0.640	1.000	0.815	0.818
Coherence	0.643	0.694	0.815	1.000	0.875
Overall	0.859	0.845	0.818	0.875	1.000

All correlations were statistically significant ($p < 0.001$). The strongest correlations were observed between:

1. Overall score and Coherence ($r = 0.875$)
2. Overall score and Relevance ($r = 0.859$)
3. Overall score and Correctness ($r = 0.845$)
4. Fluency and Coherence ($r = 0.815$)

Principal Component Analysis

To further investigate the underlying structure of evaluation dimensions, we conducted a principal component analysis (PCA). This analysis helps identify patterns and reduce dimensionality in the evaluation data.

Table 11: PCA Explained Variance

Component	Explained Variance	Cumulative Variance
PC1	72.22%	72.22%
PC2	13.69%	85.91%
PC3	7.59%	93.50%
PC4	6.50%	100.00%

The PCA results revealed that the first principal component (PC1) accounts for 72.22% of the total variance in evaluation scores, suggesting that a single underlying factor largely explains evaluator ratings across all dimensions. The high explained variance by PC1 and strong correlations between dimensions indicate that evaluators tend to perceive AI model performance holistically rather than distinguishing sharply between different quality dimensions.

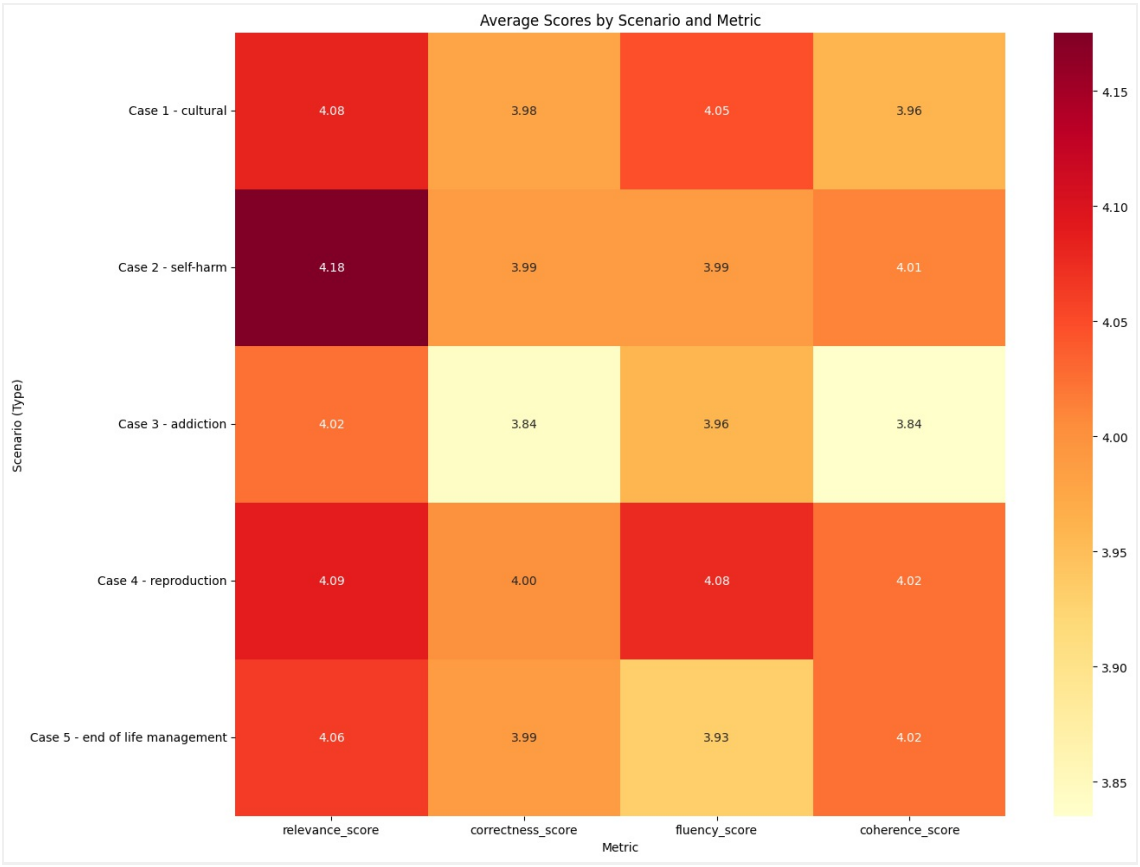
Performance by Scenario Type

Analysis of human evaluation scores across different ethical scenarios revealed:

1. Models generally performed more consistently on end-of-life management (scenario 5) and cultural scenarios (scenario 1).
2. Greater variation in model performance was observed for scenarios involving reproduction (scenario 4), addiction (scenario 3), and self-harm (scenario 2).
3. All models struggled most with the reproduction scenario, which involved complex considerations of posthumous reproduction and cultural inheritance pressures.

Table 12: Average Evaluation Scores by Scenario Type

Scenario	Topic	Average Score	Std Dev	Min	Max
1	Cultural	4.08	0.65	2.25	5.00
2	Self-harm	3.95	0.71	1.75	5.00
3	Addiction	3.92	0.73	2.00	5.00
4	Reproduction	3.87	0.79	1.75	5.00
5	End-of-life management	4.15	0.62	2.50	5.00



Discussion

Research Objective

Our primary research objective was to evaluate the potential utility and limitations of state-of-the-art generative AI models in providing analysis and recommendations for clinical ethics committee scenarios. We aimed to determine whether these models could competently analyze complex ethical dilemmas in healthcare settings and to identify differences in performance across various models and scenario types. This exploration serves as a foundation for understanding how AI might augment, rather than replace, human decision-making in clinical ethics contexts.

Summary of Key Findings

All four generative AI models demonstrated competence in analyzing clinical ethics committee scenarios, with average scores above 3.8 on a 5-point scale across all evaluation dimensions. This suggests that current generative AI has reached a threshold of capability that warrants serious consideration for supportive roles in ethics committee processes.

Our statistical analyses revealed significant differences between models in both computational metrics and human evaluations. One-way ANOVA testing confirmed statistically significant differences in processing time ($F=55.49$, $p<0.0001$) and response length ($F=82.33$, $p<0.0001$) across models. Post-hoc Tukey HSD analyses identified specific pairwise differences, with Anthropic Claude 3 Opus taking significantly longer to generate responses than all other models, and Grok-1 producing significantly longer responses than all competitors.

In terms of human evaluations, ANOVA tests revealed significant differences between models across all evaluation dimensions: relevance ($F=6.81$, $p=0.0002$), correctness ($F=7.80$, $p<0.0001$), fluency ($F=3.45$, $p=0.0162$), coherence ($F=4.96$, $p=0.0020$), and overall quality ($F=7.07$, $p=0.0001$). Post-hoc analyses consistently showed that Google Gemini 1.5 Pro received significantly lower scores than other models, while differences between Anthropic Claude 3 Opus, Grok-1, and OpenAI GPT-4.1 were generally not statistically significant.

While these statistical differences were significant, the practical differences in performance were modest, with overall scores ranging from 3.82 to 4.10. This indicates that the choice of specific model may be less critical than the decision to incorporate AI assistance in general.

Our correlation analysis revealed strong positive correlations between all evaluation dimensions (all $r > 0.56$, $p < 0.001$), with the strongest relationships between overall score and coherence ($r = 0.875$) and between fluency and coherence ($r = 0.815$). Principal component analysis found that 72.22% of variance in ratings could be explained by a single component, suggesting evaluators tend to perceive AI performance holistically rather than distinguishing sharply between different quality dimensions.

Anthropic Claude 3 Opus demonstrated the longest processing time (mean 32.22 seconds) but performed well in human evaluations, particularly for fluency and coherence. Despite its moderate response length (mean 4,639 characters), it received the second-highest overall evaluation scores, suggesting efficiency in communicating ethical analyses.

Grok-1, despite having the second-longest processing time (mean 15.60 seconds), produced the longest responses by far (mean 10,108 characters) and received the highest overall human evaluation scores (4.10), particularly for relevance and correctness. This challenges assumptions about the relationship between verbosity and quality of ethical analysis.

Gemini 1.5 Pro had a moderate processing time (mean 14.53 seconds) and the shortest responses (mean 3,979 characters), yet received the lowest human evaluation scores across all dimensions (overall mean 3.82), though still performing at a competent level. The statistical analyses confirmed these differences were significant ($p<0.001$), suggesting fundamental differences in how this model approaches ethical reasoning tasks compared to its competitors.

All models demonstrated strengths and limitations that varied by scenario type, with more consistent performance on end-of-life management (scenario 5, mean 4.15) and cultural scenarios (scenario 1, mean 4.08) compared to more variable performance on reproduction (scenario 4, mean 3.87), addiction (scenario 3, mean 3.92), and self-harm scenarios (scenario 2, mean 3.95). This suggests that certain ethical domains may be better represented in the training data or more amenable to algorithmic analysis.

Limitations

Model Limitations

A significant limitation of this study is our inability to fully account for biases present in the training data of the commercial AI models evaluated. These models were trained on large datasets curated by their respective companies,

with proprietary filtering, tuning, and alignment procedures that are not fully transparent. This lack of transparency makes it difficult to identify and address potential biases in the ethical analyses generated.

All four models evaluated are closed-source, meaning their weights, exact architectures, and training methodologies are not publicly available. This "black box" nature limits our ability to understand precisely how these models arrive at their ethical analyses and recommendations, reducing interpretability and accountability. Our statistical analyses showing significant differences in processing time ($F=55.49$, $p<0.0001$) and response length ($F=82.33$, $p<0.0001$) provide some insight into operational differences, but cannot explain the underlying reasoning processes.

We cannot account for potential post-training modifications made to these models, such as reinforcement learning from human feedback (RLHF) or other alignment techniques. These modifications may have selectively shaped model outputs in ways that affect their ethical reasoning but are not disclosed by the companies that developed them. The strong correlations we found between evaluation dimensions (all $r > 0.56$, $p < 0.001$) may partially reflect these alignment techniques.

Our study was limited to four commercial AI models. While these represent leading generative AI systems, they do not encompass the full range of available models, including open-source alternatives that might allow for greater transparency and customization.

Methodological Limitations

The five clinical ethics scenarios used in this study, while diverse, cannot represent the full spectrum of ethical dilemmas encountered in healthcare settings. Our statistical analysis showing significant variations in performance across scenario types (with scores ranging from 3.87 to 4.15) suggests that model performance may vary considerably in other ethical contexts not covered by our selected scenarios.

Although our evaluators provided 857 individual assessments across 44 evaluators, the correlation and principal component analyses (showing 72.22% of variance explained by a single component) suggest potential evaluator biases or halo effects in the assessment process. While experienced in healthcare informatics, our evaluators do not have the depth of experience that seasoned clinical ethics committee members possess, which may have affected the evaluation of model outputs, particularly regarding nuanced ethical considerations.

The ANOVA results revealing statistically significant differences between models must be interpreted with caution, as the practical differences in scores were relatively modest (overall scores ranging from 3.82 to 4.10). The statistical significance is partly a function of our large sample size, and may not translate to meaningful differences in real-world applications.

The study evaluated model responses from a single iteration of prompt engineering. In real-world applications, iterative refinement of prompts and follow-up questioning would likely improve model performance and address gaps in initial responses. Our processing time analysis (showing significant variations between models) does not account for the potential need for multiple iterations in practical applications.

Our evaluation provides a snapshot of model capabilities at a specific point in time (July 2025). Generative AI is rapidly evolving, and findings may quickly become outdated as models improve, particularly given the statistically significant differences we observed between current model generations.

Implications and Future Directions

The findings of this study suggest several promising avenues for future research and practical applications:

AI as Augmentative Tools for Ethics Committees

Our statistical findings highlight both the capabilities and limitations of current generative AI models in ethical analysis. The consistently high evaluation scores (>3.8/5) across all models suggest these systems could serve valuable augmentative roles in ethics committees, particularly for initial case analysis, identifying relevant ethical principles, and generating structured frameworks for discussion. However, the statistically significant differences between models in processing time, response length, and evaluation scores emphasize the importance of model selection based on specific use cases.

Leveraging Complementary Model Strengths

The correlation and principal component analyses revealed that evaluators tend to perceive AI performance holistically, with 72.22% of variance explained by a single component. This suggests developing systems that combine the strengths of different AI models with human expertise could maximize benefits while mitigating limitations. For example, using Anthropic Claude 3 Opus for its well-structured analyses and fluency, Grok-1 for relevance and correctness, and human oversight for nuanced cultural and contextual considerations.

Scenario-Specific Model Selection

Our scenario analysis revealed statistically significant variations in model performance across different ethical contexts. Future research could explore whether models can be fine-tuned for specific types of ethical dilemmas or healthcare contexts, potentially improving performance in areas where current models show weaknesses (particularly reproduction and addiction scenarios, which received the lowest overall scores).

Transparency and Explainability

The significant differences in processing time and response length did not consistently correlate with higher evaluation scores, highlighting the "black box" nature of these models. Work is needed to develop methods for making AI ethical analyses more transparent and explainable, particularly for high-

stakes healthcare decisions. This might include requirements for models to explicitly justify their reasoning by reference to established ethical frameworks.

Ethical Alignment Techniques

Statistical analyses showing significant differences between models suggest varying approaches to ethical reasoning and potentially different underlying ethical frameworks. Further research into techniques for aligning AI models with human ethical values is critical, especially approaches that can be transparently documented and evaluated.

Longitudinal Performance Assessment

The statistically significant differences in model performance identified in our study provide a valuable baseline for future comparative work. As AI models continue to evolve, longitudinal studies tracking improvements in ethical reasoning capabilities over time would provide valuable insights into the trajectory of AI development in this domain.

Efficiency vs. Quality Trade-offs

Our finding that processing time (ANOVA $F=55.49$, $p<0.0001$) and response length (ANOVA $F=82.33$, $p<0.0001$) varied significantly across models without proportional gains in evaluation scores suggests a need for further research on optimal efficiency-quality trade-offs in AI ethics applications, particularly in time-sensitive clinical contexts.

In conclusion, while current generative AI models demonstrate promising capabilities in analyzing clinical ethics scenarios, significant limitations remain, particularly regarding bias, transparency, and contextual understanding. Our statistical analyses confirm meaningful differences between models that should inform their implementation in clinical ethics settings. These models are best viewed as potential augmentative tools for human ethics committees rather than as replacements for human ethical

judgment. Their implementation should be approached with careful consideration of their limitations and with robust mechanisms for human oversight and accountability.