EthiQuest: LLM-Powered Ethical Questionnaire Generation for Research Review

Ishank Kapania[1], **Radhika Mamidi**[1], **Rahul Mishra**[1]
[1] Language Technologies Research Center, KCIS, IIIT Hyderabad, India
{ishank.kapania@research.iiit.ac.in, radhika.mamidi@iiit.ac.in, rahul.mishra@iiit.ac.in}

Abstract

Building upon the critical importance of ethical considerations in research, we introduce a novel task of Ethical Questionnaire Generation (EQG) for research papers. Ethical review has become an indispensable component of the research process, helping identify potential risks, biases, and societal impacts that may arise from scientific work. In this paper, we present ***EthiQuest***, a comprehensive dataset comprising 3663 research papers paired with their corresponding ethical questionnaires extracted from major conference proceedings. We explore various approaches leveraging large language models (LLMs) to automatically generate context-aware ethical questionnaires, examining the unique challenges of capturing domain-specific ethical concerns, ensuring comprehensive coverage of potential issues, and maintaining question relevance and clarity. Our experiments demonstrate the effectiveness of fine-tuned LLMs in generating pertinent ethical questions across diverse research domains. We provide detailed analysis of question quality, coverage metrics, and practical insights for deploying such systems in real-world research review processes. The EQG dataset and code can be accessed at https://github.com/Ishank-Kapania/eqg/.

Keywords: Research Ethics, Large Language Models, Ethical Review, Responsible AI

1. Introduction

Ethical evaluation has become a key part of modern research. It is an an important way to make sure that scientific discovery moves forward in a responsible way in society. It necessitates a rigorous evaluation of risks, biases, and societal consequences. Ethical review maintains scientific integrity (ALLEA European Academies, 2017), social responsibility, and beneficence through stringent examination. However, thoroughly evaluating ethical implications in research publications continues to be difficult and time-consuming.

We introduce Ethical Questionnaire Generation (EQG), which automatically generates context-aware ethical questions to probe papers' claims, assumptions, and potential oversights. EQG differs significantly from conventional NLP tasks such as text summarization or question-answering. It requires deep understanding, inference, and identifying unstated assumptions and downstream consequences. Table 1 illustrates how our method generates questions that address explicit claims (e.g., PERSUASIONFORGOOD dataset, politeness and empathy) while also formulating probing questions about secondary consent, distinctions from emotional manipulation, and missing technical safeguards.

Our approach leverages large language models trained on extensive scientific literature to identify patterns, principles, and common ethical concerns. Our goal is to generate targeted questions for papers whose methods or topics align with prior ethical issues. To support this objective, we present **EthiQuest**, a comprehensive dataset containing research papers paired with their corresponding ethical questionnaires from major conference proceedings.

The key contributions of this work are:

- To the best of our knowledge, we are the first to propose the task of Ethical Questionnaire Generation (EQG) for research papers.
- We release **EthiQuest**: 3,663 papers with corresponding ethical questionnaires.
- We explore several schemes to utilize LLMs for EQG.
- We evaluate question quality and coverage via automated, human, and LLM-driven methods.

2. Related work

Understanding ethical dimensions in scientific research is challenging (Whittlestone et al., 2019). Ethics statements are often vague, incomplete, or avoid deeper implications beyond what authors acknowledge. Recent efforts in scholarly document understanding have made significant strides. Large-scale scholarly corpora like S2ORC (Lo et al., 2020) enable broad coverage of scientific literature, while QA datasets such as QASPER (Dasigi et al., 2021) focus on evidence-based question answering over full papers. Similarly, summarization approaches—whether generating short abstracts (Cachola et al., 2020), highlighting contributions (Liu et al., 2023), or organizing content by facets (Meng et al., 2021)—aim to condense and restructure information, not to surface unstated assumptions, potential harms, or ethical trade-offs that authors may not explicitly address. The gap between

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1216–1225
11–16 May 2026.

Ethical Statement: In this work, persuasion is modelled using a publicly available dataset... utilizing the PERSUASIONFORGOOD dataset. The dataset... will be made available only with an official agreement with restriction that data would be used only for research works. The dataset is annotated... verified by our university review board. Further, persuasion is an intricate process which can be used for personal gain. Therefore, to develop a persuasive conversational AI an ethical intention must be taken into account. In this work, we choose to style persuasive dialogue generation with politeness and empathy for social work of donation to a charity organization.
Questionnaire Probing Deeper: 1. Were the original PERSUASIONFORGOOD dataset participants informed that their conversations would train AI systems to persuade others? Is secondary consent needed? 2. How do the authors distinguish between ethical "polite persuasion" and emotional manipulation, especially when the system adapts strategies based on user responses? 3. What technical safeguards prevent the model from being used for scams or predatory marketing (financial benefit)? 4. With a 67% donation success rate, should the system include mechanisms to detect and protect financially vulnerable users? 5. Is it ethical to present AI-generated empathy as genuine human concern in charity contexts? Should users be told they're interacting with an AI? 6. Could normalizing AI persuasion systems erode human autonomy in decision-making or create defensive behaviors against AI influence?

Table 1: Comparison of an ethical statement against a generated questionnaire that probes its claims and omissions. Magenta highlights concepts mentioned in the statement, while Blue indicates novel points or deeper probes raised by the questionnaire.

summarization and critical analysis becomes apparent when examining attempts to automate scientific reviewing. Systems like ReviewerGPT (Liu and Shah, 2023) and OpenReviewer (Idahl and Ahmadi, 2025) generate review-like text but lack depth for ethical critique. LimGen (Faizullah et al., 2024) generates limitations, but limitations are only one facet. Empirical studies reveal an *accountability gap* (Raji et al., 2020). The EthiCon dataset (Karamolegkou et al., 2025) catalogues ethical concerns and tries to automate this process. We need a system that reasons about what is not stated—assumptions in design (Gray et al., 2018), unconsidered populations (Birhane, 2020), and long-term deployment consequences (Brundage et al., 2018).

We address this gap by formulating **Ethical Questionnaire Generation (EQG)** as a distinct task. To our knowledge, this is the first work focused on generating targeted ethical questionnaires for research papers.

3. The EthiQuest Dataset

3.1. Dataset Collection and Construction

Researchers typically follow ethical guidelines such as ALLEA (ALLEA European Academies, 2017) and the ACM Code of Ethics,[1] and are expected to include statements addressing broader impacts.

We constructed our dataset from the ACL Anthology[2] using `scipdf_parser`[3] to segment PDFs by section. For the Ethical Questionnaire Generation (EQG) task, we extracted the main body—excluding Abstract, Introduction, Related Works, Acknowledgements, Conclusion, Appendix, References, and the Ethics Statement. This process preserves the core technical content of the paper, including its methodology, experiments, and results. The corresponding Ethics Statement served as the reference for deriving the ground-truth questionnaire. In total, our **EthiQuest** dataset comprises 3,663 peer-reviewed papers.

The dataset spans major NLP venues (ACL, EMNLP, NAACL, LREC, EACL, COLING, CONLL, TACL) from 2021–2024. Table 2 provides detailed statistics of the dataset distribution across venues and content characteristics.

3.2. Analysis of Ethical Statements

We conducted a manual analysis of 50 randomly selected articles from EthiQuest to evaluate existing practices across multiple areas.

We evaluated clarity and specificity, explicit acknowledgment of potential risks, and proposed mitigation steps. We also categorized primary ethical concerns. We found that many statements partially acknowledged risks, but a substantial portion did not, most proposed no mitigation steps for the issues they raised. As shown in Table 3, Data & Privacy was most frequent (18 papers), followed by Bias & Fairness (10). These results reveal a persistent gap: statements are often vague, omit risks, or lack concrete solutions, consistent with broader analyses (Birhane et al., 2022).

[1] https://2023.aclwcb.org/calls/ethics_guidelines/

[2] https://aclanthology.org/

[3] https://github.com/titipata/scipdf_parser/

Venue (Years Covered)	Total Count
File Counts by Venue	
ACL (2021, 2022, 2023)	960
COLING (2022)	83
CONLL (2021, 2022, 2023)	24
EACL (2023, 2024)	269
EMNLP (2021, 2022, 2023)	1051
LREC (2022, 2024)	665
NAACL (2021, 2022, 2024)	588
TACL (2021, 2022, 2023, 2024)	23
Average Statistics: Paper	
Average words	5576
Average sentences	261
Average Statistics: Ethical Statement	
Average words	165
Average sentences	8

Table 2: Dataset Statistics by Venue and Content

Category	Count
Data & Privacy	**18**
Bias & Fairness	10
Societal Impact	8
Human Subject Considerations	7
Environmental Impact	5
Other	2

Table 3: Manual analysis of ethical statements from 50 research papers.

4. Benchmark Experiments

4.1. Task formulation

This section introduces the Ethical Questionnaire Generation (EQG) task formulation. To produce ethical questionnaires for research papers, we approach the task as a Seq2Seq problem (Vaswani et al., 2017). Precisely, we craft a model designed to intake a scientific paper *R* as input and systematically produce a structured ethical questionnaire $Q = q(1:n)$, where $q(1:n)$ represents the combination of n ethical questions for *R*, sequentially generated.

4.2. Methodology

This section describes the various approaches to generate ethical questionnaires. We explore two paradigms: zero-shot generation and fine-tuning. For fine-tuning, we experiment with three distinct methods for preparing the input source text: using the full paper, employing a Dense Passage Retrieval (DPR) system, and generating an ethically-focused summary. All fine-tuning approaches use the standardized prompt structure detailed in Table 4.

Instruction:: You are a reviewer for a research paper. Generate a questionnaire that analyzes any potential ethical considerations with the practices done in the research paper.
Input:: {Input source}
Output:: {Ethical Questionnaire}
{End-of-Prompt}

Table 4: Standardized Prompt Structure for Fine-tuning.

4.3. Zero-shot

We use pre-trained LLMs without task-specific fine-tuning. Supplying full papers to medium-sized models, we prompt them to draft ethical questionnaires. This baseline lets us assess the added value of subsequent fine-tuning efforts.

4.4. Full Finetuning

We fine-tune models using the main body of each paper as input, following EthiQuest's preparation that removes auxiliary sections to retain core technical content. The paired ethical questionnaire is the target output. However, fixed context windows constrain input length, forcing truncation of longer manuscripts and potentially discarding details from later sections, which can degrade questionnaire quality.

4.5. Dense Passage Retrieval (DPR)

To mitigate context-length and information-loss in full-paper fine-tuning, we adapt Dense Passage Retrieval (Karpukhin et al., 2020) at the passage level.

Training Data Creation: The dataset is built via:

1. **Intelligent Text Chunking:** Segment each paper into semantically coherent passages using spaCy[4] for sentence boundaries and a BERT tokenizer for token counts.
2. **Semantic Matching:** For each ethical question:
   - Encode the question with a Sentence-Transformer (`all-MiniLM-L6-v2`)[5].
   - Encode all passages from the paper.
   - Compute cosine similarity between question and passage embeddings.

[4] https://spacy.io/
[5] https://www.sbert.net/

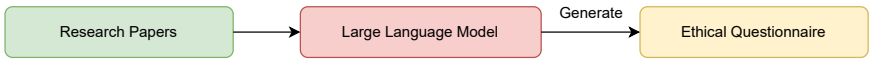

Figure 1: General architecture diagram for ethical questionnaire generation.

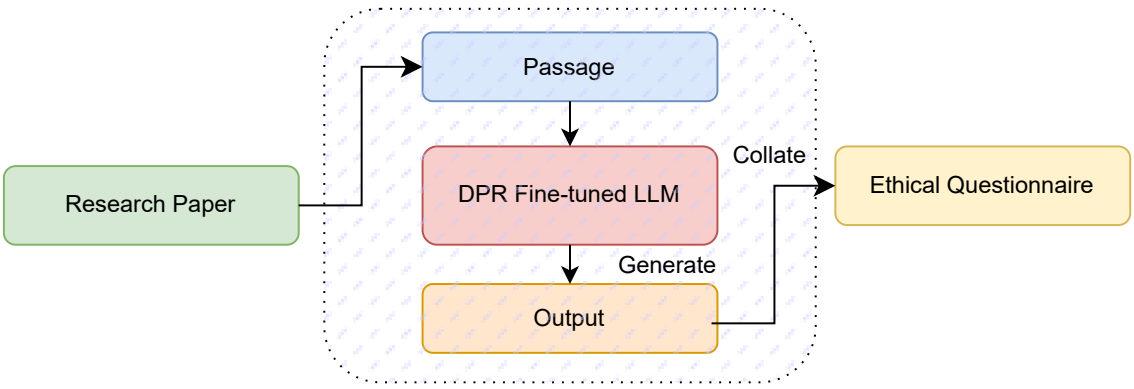

Figure 2: Architecture diagram for DPR generation.

- Select the top-3 passages with highest similarity.

3. **Dataset Construction:** Create training examples where each row contains:
   - Input: a selected relevant passage (one of the top-3).
   - Output: the matched ethical question.

Inference Pipeline: The inference operates in two phases:

1. **Passage Generation Phase:**
   - Segment the test paper using the same chunking methodology, process each passage with the fine-tuned model to propose 1–2 candidate ethical questions.
2. **Consolidation Phase:**
   - Obtain a paper-level summary and merge all passage-level questions with it.
   - Prompt a model to synthesize 7–8 final questions, deduplicated and comprehensive.

However, this approach can produce a disjointed analysis, as each passage is considered in isolation, a limitation the summary-excerpt method aims to solve.

4.6. Summary-Excerpt Approach

The majority of a paper's content does not explicitly indicate ethical issues, as revealed by manual analysis of our dataset. This understanding serves as the basis for our summary-excerpt methodology...

Training Data Creation: We employ a three-stage pipeline (detailed in Table 5) to generate an ethically-guided summary for each paper in order to generate the training data. The input source for fine-tuning is this summary. The target is the ground-truth questionnaire.

Inference Pipeline: For inference on new papers lacking an ethics statement, we adapt this pipeline to rely on the model's general knowledge as shown in Table 6.

4.7. Experimental Setup

All fine-tuning experiments were performed on various contemporary open-source Large Language Models. The Hugging Face ecosystem was utilized to construct our training pipeline. This encompassed the `transformers` (Wolf et al., 2020) and `trl`[6] libraries.

We employed Parameter-Efficient Fine-Tuning (PEFT) alongside Low-Rank Adaptation (LoRA) (Hu et al., 2021). We used Unsloth, an open-source Python library that speeds up the fine-tuning of large language models[7]. This setup made it possible to have a maximum sequence length of 30,000 tokens, which let us work with long research papers with very little truncation.

Our LoRA configuration was set with a rank (r) of 16 and an alpha of 16, targeting the attention and feed-forward network modules. Models were fine-tuned with `adamw_8bit` optimizer, a learning rate of 2×10^{-4}, and an effective batch size of 8 (2 per device with 4 gradient accumulation steps). All ex-

[6] https://github.com/huggingface/trl
[7] https://github.com/unslothai/unsloth

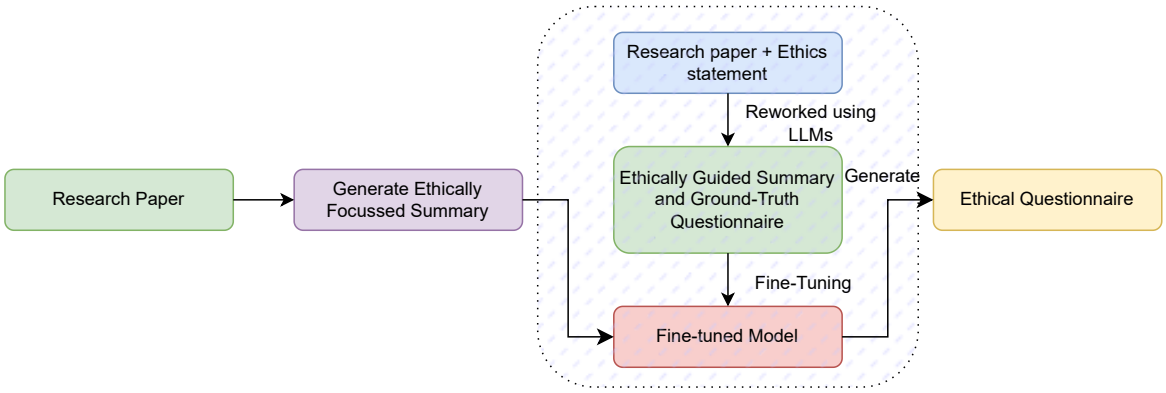

Figure 3: Architecture diagram for Summary-Excerpt Approach.

Stage	Prompt Template
Stage 1: Ethical Analysis	Here is the complete research paper: {paper_content} Here is the ethical statement for this research paper: {ethics_content} Analyze each line of the ethical statement and, for each one, identify the most relevant excerpt(s) from the research paper that support or directly relate to it.
Stage 2: Excerpt Extraction	From your previous response, please provide only the excerpts from the research paper – no ethical statement lines, no filler text, no explanations, just the raw excerpts themselves.
Stage 3: Guided Summary (Fresh Chat)	Here is a research paper: {paper_content} Here are key excerpts from this paper: {stage2_response} Please provide a comprehensive summary of this research paper, ensuring that the provided excerpts are naturally integrated into your summary. Give no filler text except the summary.

Table 5: Three-stage prompts for creating ethically-guided summaries for training using GPT-4o-mini.

Stage	Prompt Template
Stage 1: Ethical Excerpt Identification	I am a reviewer focusing on the ethical aspects of the research paper below. Based on both the contents of this paper and your prior training on research methodologies, ethical issues, and best practices from published papers, please identify and highlight specific excerpts or sections I should examine closely for potential ethical concerns or implications. Research Paper: {paper_content}
Stage 2: Excerpt Extraction	From your previous response, please provide only the excerpts from the research paper - no ethical statement lines, no filler text, no explanations, just the raw excerpts themselves.
Stage 3: Guided Summary (Fresh Chat)	Here is a research paper: {paper_content} Here are key excerpts from this paper: {stage2_response} Please provide a comprehensive summary of this research paper, ensuring that the provided excerpts are naturally integrated into your summary. Give no filler text except the summary.

Table 6: Three-stage inference pipeline using GPT-4o-mini to generate an ethically-focused summary from a new paper.

periments were performed on a single NVIDIA RTX 6000 Ada Generation GPU with 48GB of VRAM.

5. Experimental Results and Analysis

Overview. We evaluate four approaches for ethical question generation—Zero-shot, Normal Fine-

tuning, Dense Passage Retrieval (DPR), and Summary-Excerpt—on a held-out set of full research papers. We report automatic metrics (ROUGE-1/2/L (Lin, 2004) and BERTScore (Zhang et al., 2020), computed with RoBERTa-large[8]) in Table 7 for an overall model comparison, and in Table 8 to contrast the two guided pipelines (DPR vs. Summary-Excerpt). We complement these with a human evaluation of Alignment, Overlap, Actionability, and Clarity (Table 9), and an LLM-based evaluation using a TIGERSCORE-style protocol (Jiang et al., 2024) with GPT-4o[9] (Table 10). Together, these perspectives provide a balanced view of lexical overlap, semantic fidelity, practical usefulness, and error profiles.

5.1. Automatic Evaluation

Zero-shot. Zero-shot baselines produce fluent questions but tend to default to generic, broadly relevant queries. As summarized in Table 7, this appears as comparatively low bigram precision (ROUGE-2) despite reasonable semantic alignment (BERTScore). The gap highlights the difficulty of capturing paper-specific details without targeted conditioning.

Normal Fine-tuning. Supervised adaptation narrows this gap by aligning models to the task format and domain, improving bigram precision while maintaining competitive semantic similarity (Table 7). We observe stronger adherence to the desired question style and better grounding than zero-shot models.

5.2. Guided Pipelines: DPR vs. Summary-Excerpt

Dense Passage Retrieval (DPR). Using semantically chosen passages helps the model stay on topic and modestly raises bigram precision (Table 8). It surfaces ethics-relevant details and strengthens local accuracy, though it may miss cross-sectional connections.

Summary-Excerpt. The Summary-Excerpt approach yields the strongest overall automatic metrics in our experiments (Table 8). This strategy balances local evidence with full-paper context and mitigates completeness issues observed with purely passage-level conditioning.

[8] https://huggingface.co/FacebookAI/roberta-large

[9] https://openai.com/index/hello-gpt-4o

5.3. Human Evaluation

We conduct a human study on held-out papers, evaluating generated questions with a Yes/No/Partial rubric. This analysis involved four research students. Each question was assessed against four criteria: **Alignment** (Q1: Does the generated ethical question align with the research topic and methodology, or is it a generic or unrelated inquiry?), **Overlap** (Q2: Is there significant overlap b/w the generated question and known ethical statements), **Actionability** (Q3: Is the question actionable, meaning that responding to it could lead to tangible improvements in the research?), and **Clarity** (Q4: Is the ethical question clearly worded, free of ambiguity, and without redundant phrasing or grammatical errors?).

As shown in Table 9, trends from our human evaluation mirror the automatic metrics. Zero-shot systems produce highly fluent questions (tying for the top score in Clarity) but are less grounded and actionable. Guided pipelines dominate in practical usefulness: the DPR-guided model (Karpukhin et al., 2020) yields notable gains in Alignment and Actionability through passage-level grounding, while the Summary-Excerpt model delivers the best overall results.

A deeper qualitative review, guided by our human evaluation (Table 9) reveals distinct behavioral patterns. The Zero-shot model is characteristic of generating well-formed but contextually shallow questions. For instance, given a paper on a new image recognition model, it might ask, "Did the authors consider the potential for algorithmic bias?" – a valid but generic concern. In contrast, the DPR-guided model leverages specific passages to produce more grounded inquiries, such as, "The paper mentions using the CelebA dataset; were steps taken to mitigate the known gender and skin-tone biases in this specific dataset?" Showing the most advanced reasoning is the Summary-Excerpt technique. As a result, questions like, "Considering the model's high computational cost, what are the environmental effects of deploying this system on a large scale, and how does this trade off with the societal benefits mentioned at the beginning?" arise from the study's unique connections between its many sections.

5.4. LLM-based Evaluation (TIGERSCORE)

TIGERSCORE (Jiang et al., 2024) evaluates in two steps: a free-form text assessment followed by structured scoring. Using GPT-4o, penalties are applied to Completeness, Clarity, Relevance, Objectivity, Coherence, and Accuracy (lower is better, Table 10). Each aspect receives a penalty score (0.5–5) based on error severity. Results align with

Model	Approach	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Llama 3.3 70B	Zero-shot	20.5	1.7	16.6	82.3
Gemma 3 27B	Zero-shot	18.7	1.6	16.4	82.5
Llama 3.1 8B	Fine-tuning	17.9	2.3	15.2	83.3
Gemma 2 9B	Fine-tuning	19.8	**3.5**	15.3	82.1
Phi-3.5-mini-instruct	Fine-tuning	15.5	1.4	15.4	82.0
Llama 3.1 8B Instruct	Fine-tuning	20.0	1.8	16.8	83.4
Mistral-7B-v0.3	Fine-tuning	18.6	2.6	16.1	82.9
Gemma 3 12B	Fine-tuning	21.0	3.2	**18.2**	**83.5**
Phi 4 14B	Fine-tuning	**24.5**	2.7	17.2	82.0

Table 7: Performance comparison across models and approaches (automatic metrics).

Model Name	Approach	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Mistral-7B-v0.3	DPR	21.5	6.2	16.0	82.5
Llama 3.1 8B Instruct	DPR	22.0	4.5	18.6	83.7
Gemma 3 12B	DPR	25.8	7.1	17.8	83.5
Mistral-7B-v0.3	Summary-Excerpt	27.2	6.5	18.4	84.2
Llama 3.1 8B Instruct	Summary-Excerpt	26.8	**8.1**	17.9	**84.4**
Gemma 3 12B	Summary-Excerpt	**30.5**	7.3	**19.4**	83.5

Table 8: Performance by guided pipelines (DPR vs. Summary-Excerpt).

	Gemma 3 12B Fine-Tuning			Llama 3.3 70B Zero-Shot			Gemma 3 12B DPR			Llama 3.1 8B Instruct Summary-Excerpt			Gemma 3 12B Summary-Excerpt		
	Yes	No	Partial	Yes	No	Partial	Yes	No	Partial	Yes	No	Partial	Yes	No	Partial
Q1	76	10	14	24	36	40	88	4	8	86	6	8	**92**	2	6
Q2	20	30	50	8	82	10	16	24	60	10	16	**74**	12	18	70
Q3	70	12	18	30	30	40	84	6	10	82	8	10	**90**	4	6
Q4	90	4	6	**96**	2	2	92	2	6	**96**	0	4	94	2	4

Table 9: Human Evaluation of Generated Ethical Questions (N=50 papers). For Q1, Q3, and Q4 (e.g., Alignment, Actionability, Clarity), higher 'Yes' values are preferred. For Q2 (Overlap), higher 'Partial' is preferred, reflecting better capture of nuanced ethical content. All values are percentages.

earlier findings.

The DPR model scores best on Objectivity but incurs higher penalties for Completeness. The Summary-Excerpt pipeline achieves the most balanced performance, with the Gemma 3 12B variant leading on Completeness, Coherence, Accuracy, and Relevance, yielding the lowest total penalty. Normal fine-tuning remains competitive, notably achieving the best Clarity score.

5.5. Qualitative Error Analysis

A manual review of generated questions reveals several recurring error types across methods:

- **Generic but Irrelevant Inquiry:** Common in zero-shot models. Despite fluent phrasing, these questions address general ethical concerns rather than paper-specific methodologies or findings.
- **Contextual Misattribution:** DPR-guided models sometimes misattribute ethical issues discussed in related work to the current study, due to retrieved passage overlap.

6. Challenges and Future Work

Our work provides a strong foundation for Ethical Questionnaire Generation (EQG). To advance, we must address the following areas.

Inferential Reasoning Beyond Semantic Matching. A primary challenge is moving beyond surface-level pattern matching to genuine inferential reasoning. Current models can struggle with "**Contextual Misattribution**," failing to distinguish a paper's novel claims from its discussion of prior work. We need systems that detect not just what is said but what is omitted. Future work could train models adversarially to detect inconsistencies and obfuscation.

Navigating Subjectivity and Evolving Ethical Norms. Ethical standards are culturally, contextual, subjective, and responsive to the values of

	Gemma 3 12B Fine-Tuning	**Llama 3.1 8B Instruct** Summary-Excerpt	**Gemma 3 12B** DPR	**Gemma 3 12B** Summary-Excerpt
Completeness	5.50	5.70	5.80	**5.30**
Clarity	**1.24**	2.52	1.70	2.20
Relevance	3.50	3.60	2.90	**2.25**
Objectivity	1.50	**1.30**	1.70	2.25
Coherence	1.30	1.20	1.60	**1.10**
Accuracy	2.50	2.50	2.80	**2.30**
Total	15.54	16.82	16.50	**15.40**

Table 10: The table shows average penalty scores from an automatic evaluation using GPT-4o and xgptscore. For all criteria listed, a lower score indicates higher model performance. The best score in each row is highlighted in bold.

research communities (Birhane et al., 2022; Vida et al., 2023). Models trained on a singular period or environment (e.g., Western NLP conferences) risk perpetuating obsolete or constrained viewpoints (Birhane, 2020). Building systems that are capable of adapting to changing norms is a significant obstacle.

Functional Evaluation Beyond Lexical Overlap. Standard metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) are inadequate for EQG, as perceptive queries may exhibit minimal lexical overlap with the ground truth. Evaluation should use functional metrics like actionability (prompts changes in research design), specificity (grounded in paper details), and critical depth (probes unstated assumptions).

Multimodal and Code-Aware Analysis. Including figures, source code, and demos in an ethical review requires multimodality and code awareness. Architecture schematics, and code analysis shows security or fairness issues. Future systems should analyze these artifacts.

Modeling Longitudinal and Downstream Ethical Impact. Ethical effects frequently emerge gradually as artifacts are utilized in unanticipated circumstances. A major problem is figuring out how to simulate this long-term effect, such as keeping track of citation graphs and downstream uses to make a risk assessment that goes beyond the first publication.

7. Conclusion

In this paper, we introduce the novel task of Ethical Questionnaire Generation (EQG) to aid the ethical review process. To support this, we release the EthiQuest dataset, containing 3,663 papers and their corresponding questionnaires. We propose and evaluate several LLM-based methods, finding our Summary-Excerpt approach most effective for generating specific, actionable questions. A thorough evaluation using automatic, human, and LLM-based schemes validates our findings. The future work will integrate multimodal and code-aware analysis to facilitate a more comprehensive review.

8. Limitations

While EthiQuest introduces a framework for ethical Questionnaire Generation, several important limitations remain. The EthiQuest dataset is drawn solely from NLP conference proceedings (ACL Anthology), creating domain bias that limits its generalizability to biomedicine and the social sciences. EthiQuest analysis is confined to text and does not include multimodal artifacts such as figures, code, or datasets, which may have ethical consequences. The analysis is static, with no downstream or longitudinal impact tracking. While models can recognize explicit claims, they struggle to account for important omissions. Standard metrics like ROUGE are insufficient for question quality, necessitating human or LLM judgments, creating measures for insight and actionability beyond lexical overlap remains a challenge. Models based on western-centric procedures may overlook global viewpoints. Finally, EthiQuest should supplement, not replace, human ethical reasoning.

9. Ethics Statement

We adhere to responsible research practices. Data were sourced from publicly available dataset from ACL anthology, no personal or sensitive information was collected. All experiments comply with venue guidelines. We release the code under the MIT license. We release the EthiQuest dataset under the CC BY 4.0 license (same as the ACL Anthology's original license).

10. Bibliographical References

ALLEA European Academies. 2017. The European Code of Conduct for Research Integrity.

{Licensing/Redistribution:} Do ACL/publisher licenses permit large-scale parsing and redistribution of text excerpts? What is your takedown policy?
{Ground-Truth Quality:} How do you reduce bias from incomplete author ethics statements used as ground truth?
{DPR Misattribution:} What checks prevent DPR from misattributing prior-work issues to the target paper?
{Review Transparency:} If used in peer review, will authors see prompts and retrieved excerpts, and have a chance to contest the questions?
{Environmental Footprint:} What is the measured energy or carbon cost of the Summary-Excerpt pipeline, and how do you mitigate it?
{Bias Check:} Did you test the model for bias?
{Data Availability:} Is the EthiQuest dataset publicly available?

Table 11: Ethical questionnaire generated for this paper by our best-performing model (best viewed in color). Magenta indicates paper-specific probes tailored to EthiQuest, while Blue marks substandard or overly generic questions.

Abeba Birhane. 2020. Algorithmic colonization of africa. *SCRIPTed*, 17(2):389–409.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 173–184, New York, NY, USA. Association for Computing Machinery.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, S. J. Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2018. The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation. Report. February 2018.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. TLDR: Extreme summarization of scientific documents.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Abdur Rahman Bin Mohammed Faizullah, Ashok Urlana, and Rahul Mishra. 2024. LimGen: Probing the LLM's for generating suggestive limitations of research papers. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2024, Proceedings, Part II*, pages 106–124, Berlin, Heidelberg. Springer.

Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, New York, NY, USA. Association for Computing Machinery.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen(-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu) Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. NAACL 2025 System Demonstrations Track (Camera-ready version).

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yucan Lin, and Wenhu Chen. 2024. TIGERScore: Towards building explainable metric for all text generation tasks.

Antonia Karamolegkou, Sandrine Schiller Hansen, Ariadni Christopoulou, Filippos Stamatiou, Anne Lauscher, and Anders Søgaard. 2025. Ethical concern identification in NLP: A corpus of ACL anthology ethics statements. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11618–11635, Albuquerque, New Mexico. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen(-tau) Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Meng-Huan Liu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. ContributionSum: Generating disentangled contributions for scientific

papers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, New York, NY, USA. Association for Computing Machinery.

Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? an exploratory study on using large language models for paper reviewing.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: A faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089, Online. Association for Computational Linguistics.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT '20, pages 33–44, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS '17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 195–200, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Schleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Camwen Xu, Teven le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT.