Context is Key: Aligning Large Language Models with Human Moral Judgments through Retrieval-Augmented Generation

Matthew Boraske
West Chester University
West Chester, PA (USA)

Richard Burns
West Chester University
West Chester, PA (USA)
rburns@wcupa.edu

Abstract

In this paper, we investigate whether pre-trained large language models (LLMs) can align with human moral judgments on a dataset of approximately fifty thousand interpersonal conflicts from the AITA (Am I the A******) subreddit, an online forum where users evaluate the morality of others. We introduce a retrieval-augmented generation (RAG) approach that uses pre-trained LLMs as core components. After collecting conflict posts from AITA and embedding them in a vector database, the RAG agent retrieves the most relevant posts for each new query. Then, these are used sequentially as context to gradually refine the LLM's judgment, providing adaptability without having to undergo costly fine-tuning. Using OpenAI's GPT-4o, our agent outperforms directly prompting the LLM while achieving 83% accuracy and a Matthews correlation coefficient of 0.469 while also reducing the rate of toxic responses from 22.53% to virtually zero. These findings indicate that the integration of LLMs into RAG agents is an effective method to improve their alignment with human moral judgments while mitigating toxic language.

Introduction

The rapid advancement of Large Language Models (LLMs) has revolutionized natural language processing (NLP). However, while these models excel at tasks such as machine translation and question answering (Brown et al. 2020; Raffel et al. 2020), their ability to evaluate the morality of text remains largely unexplored. Morality is difficult to quantify due to its nebulous nature, as it is uniquely defined by individuals based on factors such as their socioeconomic status, cultural background, and human experiences (Yates and De Oliveira 2016; Iyer, Weinberg, and Bagot 2022).

We present an investigation into utilizing LLMs to evaluate morality through the lens of a task completed on the AmITheA******" (AITA) subreddit, an online forum hosted by Reddit with over twenty-three million subscribers [1]. In this work, we treat morality judgments as discrete labels determined by large crowd-sourced consensus on the AITA subreddit. Users submit posts that describe one of their interpersonal conflicts and others comment with one of four judgments on the morality of the submitting user: (1) NTA not the a******, (2) YTA you are the a******, (3) NAH no a****** here, and (4) ESH everyone sucks here. An optional justification for the judgment is also usually provided. Reddit promotes the most relevant comments through a community voting system, and the submission receives a label morality judgment with the top-ranking comment.

Despite the fact that LLMs excel in question-answering tasks, the moral judgment task presented on AITA, where the true answer depends on a varying sense of morality aggregated from a large group, represents a unique challenge. Unlike factual tasks, moral judgments often hinge on precedent or analogous scenarios (Hofmann et al. 2014), making the retrieval of similar ethical dilemmas especially valuable. We show that by using retrieval-augmented generation (RAG) to give the LLM access to examples of AITA judgments, the LLM can be steered to align its moral judgments with the human subscribers. RAG combines external document retrieval with LLM response generation by first retrieving relevant documents from a knowledge base and then incorporating them into the prompt to inform the response (Lewis et al. 2020).

To develop a RAG agent for the morality judgment task presented on the AITA subreddit, we collected a dataset of nearly 50,000 submissions and their top-ranking comment pairs from 2018 to 2022. We then embedded these using OpenAI's text-embedding-3-large model (OpenAI 2024b) and stored them in a vector database, which serves as the knowledge base for the RAG agent. When making a moral judgment on a new submission, the agent first embeds its text and retrieves the five most similar submissions in the knowledge base. It then sequentially provides these submissions and the AITA judgments in their top comments as context, allowing the LLM to iteratively refine its moral judgment based on these precedents.

Using GPT-4o (OpenAI 2024a) as the LLM for our RAG agent, we achieved 83% accuracy and a Matthews correlation coefficient of 0.469, representing absolute improvements of 6% and 0.102 over our direct LLM prompting experiments. These gains are due to the RAG agent using the provided context to better distinguish the moral boundaries between AITA scenarios, particularly for NAH and ESH

[1] https://www.reddit.com/r/AmIthe*******/

judgments, where a single party is not at fault and there is a lesser degree of consensus on the appropriate verdict.

In light of these advancements, it is important to note that online forums can frequently contain toxic language, which poses a significant risk since models trained on these sources can learn to perpetuate harmful norms. By examining toxicity alongside moral judgments, we explore how a retrieval-based approach can ensure that our agent not only aligns with community verdicts but also mitigates harmful outputs, thereby maintaining respectful and safe responses.

We believe that our work is the first to explore building an RAG agent to directly model community-driven moral judgments at scale by leveraging a large corpus of user-submitted ethical dilemmas. Although existing retrieval-based methods have shown success in factual tasks such as knowledge-intensive question answering, our approach extends these methods to the domain of subjective moral reasoning.

The remainder of this paper is structured as follows. First, we review the literature on AI ethics and moral reasoning, retrieval-augmented generation, and the utilization of social media data for training language models. Next, we detail our methodology, focusing on two key components: the creation of the AITA dataset and the design of our RAG agent. Finally, we present our experimental results, discuss key findings, and conclude with directions for future research.

Literature Review

AI Ethics and Moral Reasoning

Seminal discussions on AI ethics contrast top-down (rule-based) and bottom-up (learning-based) approaches to imbue systems with moral reasoning (Moor 2006; Anderson and Anderson 2007). Wallach et al. present a comprehensive approach through the LIDA cognitive architecture, which decomposes the moral decision-making task into emotional and rational components (Wallach, Franklin, and Allen 2010). Meanwhile, Jiang et al. propose Delphi, a framework that trains deep neural networks to directly reason about ethical judgments (Jiang et al. 2021).

Lourie et al. underscore the importance of accounting for moral ambiguity in AI systems through their SCRUPLES dataset, which reveals how ethical judgments on real-life situations can lack consensus (Lourie, Le Bras, and Choi 2021). Tangentially, Alhassan et al. propose fine-tuning large language models (LLMs) on the AITA subreddit, demonstrating high accuracy on moral judgments when reducing the task to binary classification between NTA and YTA scenarios. In contrast, our work addresses the multi-class complexity introduced by the more nuanced NAH and ESH judgments, where blame is shared or absent (Alhassan, Zhang, and Schlegel 2022).

RAG to Enhance AI Decision-Making

Recent developments in Retrieval-Augmented Generation (RAG) offer promising methods to ground LLM output in external knowledge (Lewis et al. 2020; Izacard and Grave 2020). By embedding data sources in a vector database, retrieving the most relevant passages for a query, and feeding these passages alongside the query to the LLM, RAG reduces hallucinations and aligns model output with factual or contextual references. RAG agents have been used mainly for domain-specific tasks, such as medical advice (Raja, Yuvaraajan, and others 2024; Sree et al. 2024) or scientific research (Lála et al. 2023), and this principle can be extended to moral judgments by retrieving ethically similar precedents from a repository of real-world conflicts.

Usage of Social Media Text as Training Data for Language Models

Social networks have become rich sources of spontaneous discourse that reflects daily moral and ethical values. Popular uses of social media data include training models for hate speech detection, on specific sites such as Twitter (Watanabe, Bouazizi, and Ohtsuki 2018) or more general platforms (Hartvigsen et al. 2022), as well as modeling user profiles (Jiang and Ferrara 2023). In particular, the AITA subreddit has been used for moral classification because it offers crowd-labeled judgments from everyday users on interpersonal conflicts (Alhassan, Zhang, and Schlegel 2022).

However, these data come with inherent challenges: they often exhibit noise and toxicity, as highlighted by prior stance-detection and controversy prediction studies. For example, Sadeque et al. (Sadeque et al. 2019) examined how the social media stance on atheism-skepticism debates featured heavy polarization and incivility, while Hessel and Lee (Hessel and Lee 2019) investigated early indicators of controversy in online forums, finding that morally charged posts tend to elicit rapid, polarized reactions. Together, these studies emphasize the importance of techniques, such as RAG, to handle diverse moral viewpoints and avoid oversimplification when training models on social media text.

Methodology

In this section, we outline our methodology to generate moral judgments on AITA submissions using a retrieval-augmented generation (RAG) agent. Our approach consists of two key components: (1) the construction of the AITA dataset, including data collection, preprocessing, and embedding in a vector database, and (2) the design and implementation of the RAG agent. By combining these elements, our agent leverages historical AITA posts to ground the reasoning of the LLM so that generated judgments align more closely with human perceptions of morality.

AITA Dataset

Our AITA dataset[2] was created from two data files containing 1.7 million submissions and 72.1 million comments made from 2018 to 2022[3]. On Reddit, users cast upvotes or downvotes on both submissions and comments, each vote increasing the score by one. Thus, the final score reflects the net effect of all positive and negative votes. We leveraged

[2]The AITA dataset is available at [link omitted for blind review]

[3]The files used to build the AITA dataset are available at `http://bit.ly/42jJ7aq` and were made available for academic research by Reddit user Watchful1

Moral Judgment	Number of Samples	Inter-Annotator Agreement (%)	Toxicity Rate (%)
NTA (*not the a****)**	40,549 (82.18%)	89.16	22.83
YTA (*you're the a****)**	5,576 (11.30%)	82.46	19.64
NAH (*no a***** here*)**	1,887 (3.82%)	35.60	9.17
ESH (*everyone sucks here*)	1,331 (2.70%)	37.43	30.82
All	49,343 (100%)	84.03	22.93

Table 1: Distribution, inter-annotator agreement, and toxicity rate, in our AITA dataset.

these scores when scraping submissions and their top ten comments to apply the following filters:

1. Limited samples to those posted between the years 2018 and 2022.
2. Removed submissions with a total score below fifty, which we considered the threshold for meaningful positive engagement.
3. Removed submissions where the top comment had a score below ten, under the assumption that such a threshold reflects a broader community acceptance of the moral judgment.
4. Removed any submissions where the top comment did not follow the style convention promoted on the AITA subreddit, which is an AITA classification immediately followed by a justification. This structure is critical to our RAG agent design, as it ensures that each retrieved document includes a proper moral judgment for the described conflict.
5. Removed all other top ten comments that had a score below ten.

Table 1 summarizes the distribution, inter-annotator agreement, and toxicity rates for each moral judgment in our AITA dataset. The samples are highly skewed toward *not the a******* (NTA), representing 82.18% of submissions. In contrast, *you are the a******* (YTA) comprises 11.30%, while *no a****** here* (NAH) and *everyone sucks here* (ESH) occur even less frequently (3.82% and 2.70%, respectively). One possible reason for this could be that people who submit conflicts to AITA typically seek validation that their behavior was moral, resulting in a NTA judgment. Because our objective is to model real-world behavior as faithfully as possible, we chose to retain this natural class imbalance, leaving any exploration of synthetic data augmentation or other balancing techniques for future work.

We define inter-annotator agreement as the percentage of the second through tenth top comments that share the same verdict as the first top comment. As seen in Table 1, NAH (35.60%) and ESH (37.43%) have notably lower agreement compared to NTA (89.16%) and YTA (82.46%), suggesting that multiple perspectives are more common in shared-fault or no-fault scenarios.

To examine how frequently AITA comments contain harmful language, we quantified toxicity for each moral judgment by using a RoBERTa model[4] fine-tuned on the Toxigen dataset (Hartvigsen et al. 2022). This model outputs whether text is toxic or benign. Overall, the AITA dataset has a substantial average toxicity rate of 22.93%. Toxicity is lowest (9.17%) in NAH conflicts and highest (30.82%) in ESH conflicts. For a broader context of how AITA toxicity rates compare with other social media datasets, Founta et al. (Founta et al. 2018) analyzed 80,000 Twitter posts and found that 18.5% contained toxic or inappropriate content.

RAG Agent Design

Figure 1 illustrates the workflow of our retrieval-augmented generation (RAG) agent, which leverages historical AITA submissions to inform and refine its moral judgments. Below, we outline the key steps in the workflow and discuss the rationale behind our design choices, particularly our decision to provide retrieved documents to the large language model (LLM) in a sequential rather than all-at-once manner.

1. *Knowledge Base Creation*: We partitioned our AITA dataset into 80/20 splits: a training set to act as the agent's knowledge base and a test set used for evaluation. We vectorized each textual sample in the training set (i.e., each conflict and its moral judgment) using OpenAI's text-embedding-3-large embedding model. We stored these embeddings, along with the original text, in a vector database that supports similarity search.
2. *Embedding the New Conflict*: When our agent receives a new AITA conflict from the test set, it embeds the text with the same embedding model used for the training set. This ensures consistency in how the vector representations are generated and enables subsequent similarity retrieval.
3. *Retrieving Similar Conflicts*: A cosine vector similarity search is completed between the newly embedded conflict and those stored in the vector store to find the five most similar stored conflicts. We chose to limit retrieval to five conflicts to provide a sufficiently diverse set of precedents for the LLM while keeping computational overhead manageable. These prior conflicts serve as precedents for how related moral dilemmas were judged and justified in the AITA community.

[4] https://huggingface.co/tomh/toxigen_roberta

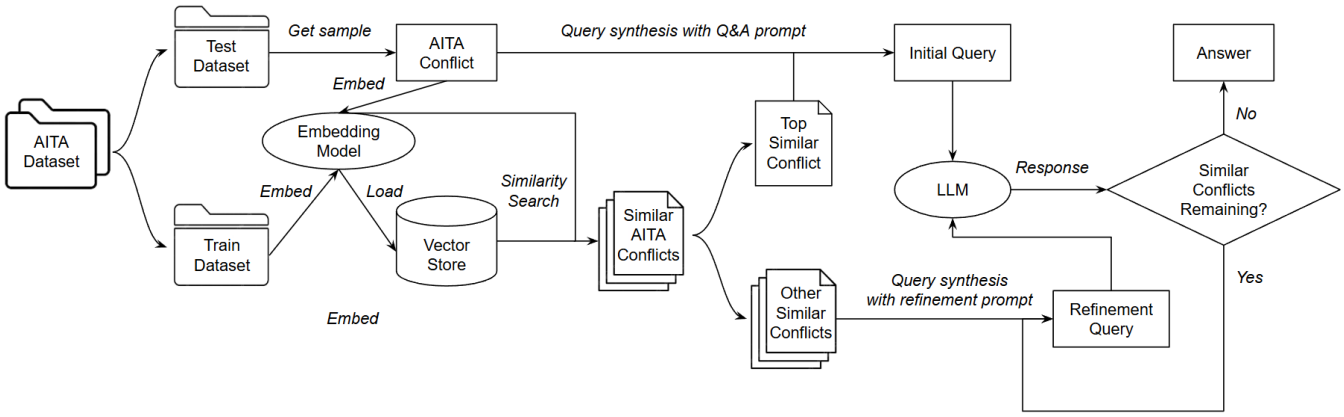


Figure 1: RAG workflow for AITA response generation. The agent retrieves a set of similar submissions, uses the highest-ranked one as context when generating the initial response, then iteratively refines using the remaining retrieved submissions.

4. *Reranking of Retrieved Conflicts*: The retrieved conflicts are reranked using Cohere's rerank-english-v3.0 model(Shi and Reimers 2024), which re-evaluates each conflict's relevance against the query. This provides a more nuanced assessment than simple vector distance alone, ensuring that the most semantically aligned examples are placed at the top of the retrieval set (Pradeep, Sharifymoghaddam, and Lin 2023). By prioritizing higher-quality precedents, our agent can ground its moral judgments in the conflicts most closely mirroring the user's situation.
5. *Initial Query Synthesis and Response*: From the list of retrieved conflicts, our agent selects the top-ranked entry and synthesizes an initial query for the LLM that consists of the text of the new conflict and the top similar conflict, including its AITA classification and supporting rationale. The LLM component then generates an initial response that contains a moral judgment and justification.
6. *Sequential Refinement*: Our agent selects the next-highest ranked entry from the retrieval set and prompts the LLM to refine its moral judgement using the selected entry. This process continues until all retrieved conflicts have been considered and the final answer is returned.

   The prompt templates our agent uses for both the initial query and refinement process are detailed in our project's codebase documentation.(Citation omitted for blind review.) We chose to sequentially provide context, rather than prompting the LLM with all similar conflicts at the same time, for two primary reasons:

   (a) *Cognitive Load Management*: Modern LLMs including GPT-4o have a finite context window (Brown et al. 2020; Achiam et al. 2023), echoing the human limits on working memory (Miller 1956). Consequently, providing all similar conflicts in one prompt can overwhelm the model, thereby reducing overall output quality.

   (b) *Iterative Reasoning*: When given all information at once, the model may simply try to compress different details. Inspired by methods such as chain-of-thought prompting (Wei et al. 2022) and iterative self-improvement (Zelikman et al. 2022), our approach sequentially introduces new context, guiding the model to update or refine its moral judgment incrementally.

Results

We implemented and evaluated two agents on our held-out test set (9,867 samples):

1. **Basic Agent - Baseline**: Single direct prompt to GPT-4o
2. **RAG Agent**: Utilizing retrieval-augmentation and providing GPT-4o with similar AITA posts as sequential context

Classification Performance

Table 2 reports the overall classification metrics - precision, recall, F1, accuracy, and Matthews correlation coefficient (MCC) - for the two approaches. The RAG Agent outperforms the Basic Agent on every metric, improving accuracy from 77% to 84% and improving MCC from 0.357 to 0.469. These gains indicate that retrieving analogous examples can better ground GPT-4o's judgments, leading to stronger alignment with the AITA community.

Agent	Prec.	Recall	F1	Acc.	MCC
Basic	0.79	0.77	0.77	0.77	0.357
RAG	0.84	0.83	0.84	0.84	0.469

Table 2: Classification metrics for all AITA judgments.

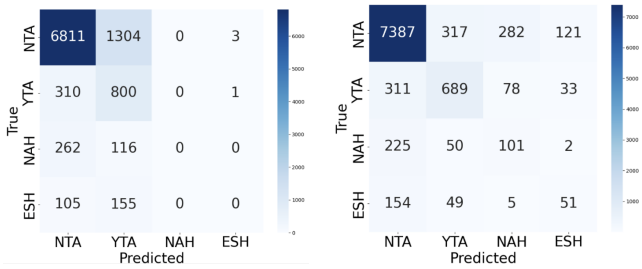
Figure 2 presents confusion matrices for the Basic and RAG Agents. For the Basic Agent, all predicted moral judgments are NTA or YTA and the model never predicts NAH or ESH, meaning that all samples of those classes are misclassified. This explains the zero precision, recall, and F1 scores for NAH and ESH (see Table 1).

Compared to the Basic Agent model, the RAG Agent does classify using all four judgments. Although the majority class (NTA) still accounts for the majority of the predictions, we see nontrivial portions of NAH and ESH correctly

	NTA			YTA			NAH			ESH		
Agent	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Basic	0.91	0.84	0.87	0.34	0.72	0.46	0.00	0.00	0.00	0.00	0.00	0.00
RAG	0.91	0.91	0.91	0.62	0.62	0.62	0.22	0.27	0.22	0.25	0.20	0.22

Table 3: Classification metrics for NTA, YTA, NAH, and ESH Judgments

identified. Retrieving and utilizing relevant examples helps the RAG Agent handle more nuanced moral dilemmas involving shared blame that were much less prevalent in the training data.



(a) Directly Prompting GPT-4o (b) RAG Agent

Figure 2: Confusion Matrices for AITA Judgements

Breakdown by Moral Judgment Table 3 summarizes the precision, recall, and F1 score by class. Both approaches perform strongly on the dominant NTA label, but the RAG agent increases the recall from 0.84 to 0.91 while keeping the precision at 0.91. The greatest improvement are in the minority classes, where the F1 measure of YTA increases from 0.46 to 0.62, and the F1 scores of ESH and NAH both increase from 0.00 to 0.22. The retrieval augmentation approach of the RAG Agent improves model performance on a per-judgment basis, especially for non-single-party-fault cases (NAH) and mutual-blame scenarios (ESH).

Quality of Justifications

Beyond predicting moral judgments, we evaluated both the toxicity and the textual similarity of the generated justifications, as reported in Table 4. To measure toxicity, we continue to use a RoBERTa model fine-tuned on the Toxigen dataset (Hartvigsen et al. 2022). Our Basic and RAG agents produced remarkably low toxicity rates (0.0004), a stark contrast to the toxicity rates of 22.93% and 22.53% observed in training and test sets, respectively. We attribute this improvement to the internal safety mechanisms of GPT-4o and additional prompt constraints, which we handcrafted to effectively mitigate harmful or profane output. Although these measures ensure that responses remain civil, they inevitably sanitize the often blunt or combative tone typical of authentic AITA discussions. Balancing authenticity with ethical considerations that prevent harm remains an open question for future work.

We also evaluated how closely the justifications match the reference comments and how verbose these justifications tend to be. To gauge text similarity, we measured ROUGE-Lsum, which was slightly higher for the Basic Agent but remains competitive for our RAG Agent. This is likely due to the increase in length ratio, defined as the average number of tokens in each generated justification divided by the number of tokens in the corresponding reference comment. While the increased length ratio indicates that the retrieved context encourages more detailed justifications, a large increase in length relative to the reference decreases ROUGE-Lsum because more words reduces the fraction of shared text.

Agent	**Toxicity rate**	**ROUGE-Lsum**	**Length Ratio**
Basic	0.0003	0.145	1.84
RAG	0.0004	0.119	2.96

Table 4: Justification Toxicity and Similarity

Discussion

In this section, we discuss the implications of our findings and situate them within the broader context of moral reasoning in artificial intelligence.

RAG as a Form of Case-Based Moral Reasoning

A key finding in our study is that leveraging RAG can substantially improve the alignment of LLMs with human moral judgments. This benefit can be understood as a type of case-based reasoning, where analogical thinking is central (Kolodner 2014). By retrieving AITA posts alongside their verdicts, the RAG agent emulates how people recall similar past experiences when forming moral judgments. Rather than relying solely on a single prompt to capture all context, the model draws on concrete precedents that reflect the collective moral standards of the AITA community.

This case-based paradigm is particularly beneficial when evaluating the less common and more controversial verdicts of NAH and ESH. These labels typically require a finer-grained understanding of the conflict, such as shared or absent blame. Our baseline approach of directly prompting GPT-4o never resulted in an NAH or ESH judgment because it collapsed the judgments into NTA or YTA. By comparing new scenarios with multiple past conflicts, our RAG Agent more readily detects that some conflicts do not involve clear fault (NAH) or feature mutual responsibility (ESH). In doing so, retrieval provides the model with contextual cues that encourage more nuanced classifications. Ultimately, this approach offers a structured and flexible strategy for aligning LLM output with the complexity and diversity of authentic human perspectives.

Implications of Toxicity

Although aggressively filtering out harmful language reduces the immediate risk of perpetuating offensive discourse, it can also dilute the emotional tone found in moral conflicts. In our work, we retained all toxic language in the training corpus to preserve the data's original intent, instead relying on a combination of GPT-4o's built-in safeguards and our prompts to enforce safe responses. While our approach proved effective at mitigating toxicity, the "black-box" nature of proprietary LLMs like GPT-4o highlights the need for further research on balancing harm reduction with the preservation of authentic user expression.

Managing Class Imbalance

One of the central obstacles in the AITA dataset is the substantial skew toward the NTA judgments, with the minority classes YTA, NAH, and ESH collectively comprising less than 20% of the data. We chose to maintain this imbalance to faithfully reflect the natural distribution of AITA conflicts, as artificially modifying that distribution can potentially alter the model's understanding of real-world user behavior. Indeed, approaches such as oversampling or generating synthetic examples could introduce new biases or cause the model to overestimate the prevalence of rarer labels in actual AITA posts. However, there are several avenues to mitigate the impact of the imbalanced dataset, which are future work:

1. **Synthetic Data Generation:** Using an external LLM to create new samples that emulate the style and content of real submissions could help expand the NAH and ESH classes (Li et al. 2023). Although this method has the benefit of increasing the coverage of the minority class, it must be approached with caution to avoid distorting the genuine moral norms of the community.
2. **Weighted Sampling:** In retrieval-based systems, adjusting the probabilities of sampling or assigning higher weights to minority classes can help ensure that the agent retrieves and observes a balanced range of judgments (He and Garcia 2009). Such weighting could encourage the model to consider more examples of ESH or NAH when forming responses, counteracting any bias stemming from the majority class.

Limitations

Although these results demonstrate the promise of RAG in aligning LLMs with human morality, they are subject to several limitations:

1. The substantial class imbalance (with a majority of posts labeled NTA) may bias both retrieval and generation.
2. While sequential feeding contextual examples to the LLM prevents exceeding the LLM context window (128k tokens for GPT-4o (OpenAI 2024a)), it introduces the possibility that key context is overlooked if it appears in later examples.

Conclusion

In this paper, we investigated whether pre-trained LLMs can align with the community-based moral judgments found on the AITA subreddit. By providing access to historical AITA posts via a retrieval-augmented generation (RAG) pipeline, we demonstrated that GPT-4o can more accurately and consistently match consensus verdicts, particularly for nuanced or less common scenarios, than when directly prompted, as represented by our Basic Agent. Empirically, our RAG agent achieved 83% accuracy and a Matthews correlation coefficient of 0.469, notably improving over direct prompting; furthermore, our approach underscores how RAG can be used to guide LLMs on how to strike the right balance between maintaining authenticity and minimizing harmful language. These gains highlight the value of "case-based" moral reasoning, in which the retrieved examples serve as precedents to guide the classification.

Our findings are underscored by several important considerations that warrant future work. First, the class imbalance in AITA posts, with a strong bias towards the NTA judgment, poses both methodological and ethical challenges, as minority judgments are correspondingly rarer. Although the RAG approach better handles these minorities, additional techniques, such as synthetic data generation or weighted sampling, could further address imbalances. Second, the AITA community on Reddit reflects a specific cultural and social environment, limiting generalization to general human morality. Future studies might explore how retrieval-augmented methods perform on data drawn from broader cultural contexts. Finally, although our agent retrieves similar cases and provides them to the LLM sequentially, further improvements could be achieved by exploring different retrieval strategies, varying how the retrieved context is presented to the LLM, and investigating alternative embeddings, language models, and reranking heuristics.

Overall, this work demonstrates that retrieval-augmented LLMs offer a scalable way to align AI outputs with human moral judgments, without the need for extensive fine-tuning or architectural modifications. Our approach not only provides a strong baseline for case-based AI moral reasoning but also serves as a foundation for future research aimed at producing AI systems that respect human morality.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alhassan, A.; Zhang, J.; and Schlegel, V. 2022. 'am i the bad one'? predicting the moral judgement of the crowd using pre–trained language models. In *Proceedings of the thirteenth language resources and evaluation conference*, 267–276.

Anderson, M., and Anderson, S. L. 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine* 28(4):15–15.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21(9):1263–1284.

Hessel, J., and Lee, L. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. *arXiv preprint arXiv:1904.07372*.

Hofmann, W.; Wisneski, D. C.; Brandt, M. J.; and Skitka, L. J. 2014. Morality in everyday life. *Science* 345(6202):1340–1343.

Iyer, E. S.; Weinberg, A.; and Bagot, R. C. 2022. Ambiguity and conflict: Dissecting uncertainty in decision-making. *Behavioral Neuroscience* 136(1):1.

Izacard, G., and Grave, E. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Jiang, J., and Ferrara, E. 2023. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893*.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Kolodner, J. 2014. *Case-based reasoning*. Morgan Kaufmann.

Lála, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodriques, S. G.; and White, A. D. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

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

Li, Z.; Zhu, H.; Lu, Z.; and Yin, M. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Lourie, N.; Le Bras, R.; and Choi, Y. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13470–13479.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63(2):81.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21(4):18–21.

OpenAI. 2024a. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2025-01-24.

OpenAI. 2024b. New embedding models and api updates. `https://openai.com/index/new-embedding-models-and-api-updates/`. Accessed: 2025-01-24.

Pradeep, R.; Sharifymoghaddam, S.; and Lin, J. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21(140):1–67.

Raja, M.; Yuvaraajan, E.; et al. 2024. A rag-based medical assistant especially for infectious diseases. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, 1128–1133. IEEE.

Sadeque, F.; Rains, S.; Shmargad, Y.; Kenski, K.; Coe, K.; and Bethard, S. 2019. Incivility detection in online comments. In *Proceedings of the eighth joint conference on lexical and computational semantics (* SEM 2019)*, 283–291.

Shi, S., and Reimers, N. 2024. Introducing Rerank 3: A New Foundation Model for Efficient Enterprise Search & Retrieval.

Sree, Y. B.; Sathvik, A.; Akshit, D. S. H.; Kumar, O.; and Rao, B. S. P. 2024. Retrieval-augmented generation based large language model chatbot for improving diagnosis for physical and mental health. In *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, 1–8. IEEE.

Wallach, W.; Franklin, S.; and Allen, C. 2010. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in cognitive science* 2(3):454–485.

Watanabe, H.; Bouazizi, M.; and Ohtsuki, T. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access* 6:13825–13835.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35:24824–24837.

Yates, J. F., and De Oliveira, S. 2016. Culture and decision making. *Organizational behavior and human decision processes* 136:106–118.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems* 35:15476–15488.

Appendix A: AITA Prompt Templates

AITA Background Information:

AITA (Am I The A**hole) is a format where people share personal conflicts and ask for judgment.
In these situations:
- The writer describes a specific conflict or dilemma they're involved in.
- They explain their actions and the actions of other involved parties.
- They share relevant context like relationships, history, and circumstances.
- The core question is always about who's at fault in the conflict.
- Judgments focus on actions and choices, not on judging people as individuals.

AITA Prompt Rules:

You MUST follow these rules
- Your AITA classification choices are limited to you're the a**hole (YTA) when the writer is causing the conflict, not the a**hole (NTA) when someone other than the writer is causing the conflict, no a**holes here (NAH) when no one is causing the conflict, and everyone sucks here (ESH) when everyone is causing the conflict.
- You MUST maintain consistency with the reasoning in the provided context and are NOT allowed to make independent moral judgments.
- You are NOT allowed to cite the specific details from the context examples to support your judgement.
- Use the context ONLY as a template for writing in a similar style (phrasing and expressions) and tone (casual vs formal), identifying the right amount of detail in justifications, and presenting your verdict strictly in the form of the AITA classification followed by the justification.
- Do NOT use racist, derogatory, or explicit language even if its used in the judgement example.

Initial Judgment Prompt Template:

Include AITA BACKGROUND INFORMATION
Study this NEW AITA conflict and judgement carefully:
`{ context string }`
Using the above as context, provide your initial judgment of this ORIGINAL AITA conflict:
`{ query string }`
You MUST follow these rules:
- You MUST judge the ORIGINAL AITA conflict and NOT the NEW conflict."

Include AITA PROMPT RULES
Your initial judgment:

Refinement Prompt Template:

Include AITA BACKGROUND INFORMATION
You are refining your judgment on this ORIGINAL AITA conflict:
`{ query string }`
This is your previous judgment of the ORIGINAL AITA conflict:
`{ existing answer }`
Now study this potentially SIMILAR AITA conflict and judgement:
`{ context message }`
Using this as context, focus EXCLUSIVELY on refining your judgment of the ORIGINAL AITA conflict by following these rules:
- If the SIMILAR conflict reveals stronger reasoning for a different classification: Change your judgment.
- If the SIMILAR conflict provides additional support for your current classification: Enhance your reasoning.
- If the SIMILAR conflict offers new perspectives: Incorporate them regardless of whether they change or support your classification.
- If the SIMILAR conflict seems less relevant or compelling: Maintain your current judgment. You also MUST continue to follow these rules:

Include AITA PROMPT RULES

Table 5: Iterative Prompt Templates Used by Our RAG Agent to Construct LLM Queries