# Integrating Emotional and Linguistic Models for Ethical Compliance in Large Language Models

**Anonymous ACL submission**

## Abstract

This research develops advanced methodologies for Large Language Models (LLMs) to better manage linguistic behaviors related to emotions and ethics. We introduce DIKE, a framework that enhances the LLMs' ability to internalize and reflect universal human values, adapting to varied cultural contexts to promote transparency and trust among users. The methodology involves detailed modeling of emotions, classification of linguistic behaviors, and implementation of ethical guardrails. Our innovative approaches include mapping emotions and behaviors using self-supervised learning techniques, refining these guardrails through adversarial reviews, and systematically adjusting outputs to ensure ethical alignment. This framework establishes a robust foundation for AI systems to operate with ethical integrity and cultural sensitivity, paving the way for more responsible and context-aware AI interactions.

## 1 Introduction

Recent papers by (Bengio et al., 2024) and (Dalrymple et al., 2024) underscore the importance of addressing AI risks and safety concerns given the technology's rapid advancement. This research introduces an alternative to Reinforcement Learning from Human Feedback (RLHF) (OpenAI, 2023; Ouyang et al., 2022) to address ethical concerns in Large Language Models (LLMs). While RLHF has demonstrated success, it faces notable challenges. First, it is prone to biases inherent in human feedback, exacerbated by today's increasingly polarized society. Second, it is susceptible to reward hacking (Christiano et al., 2023; Skalse et al., 2022), potentially leading LLMs to adopt unethical or harmful behaviors. Third, RLHF has been reported to degrade the performance of ChatGPT due to the "forgetting effect," as demonstrated by (Vianna et al., 2023; Kirkpatrick et al., 2017).

A significant limitation of current research is its narrow focus on suppressing specific, undesirable behaviors, such as movie ratings or toxic language. This "Whack-A-Mole" approach rarely addresses the underlying causes and can lead to unintended consequences, like the aforementioned "forgetting effect." Fixing one issue in an LLM may inadvertently worsen others, much like how addressing a surface-level addiction problem can sometimes reveal deeper issues and trigger side effects in humans (Sinha, 2008; Torrens et al., 2005). Another limitation is the "one-size-fits-all" nature of many RLHF implementations, which fail to adapt to the diverse cultures and values of different users, as noted by (Dalrymple et al., 2024).

To overcome these limitations, we introduce DIKE, a framework independent of the LLM itself. Standing for **D**iagnostics, **I**nterpretation, **K**nowledge-independent learning, and **E**thical guardrails, DIKE is named after the Greek goddess of justice and moral order. It aims to enhance ethical compliance in LLMs through transparent, interpretable, and independent oversight mechanisms. Functioning as a separate behavioral advisor, DIKE evaluates and guides the LLM's responses based on established ethical standards, without modifying the underlying neural structures or parameters. This architectural separation ensures that ethical enhancements do not compromise the LLM's ability to represent knowledge accurately.

To achieve adaptability and cultural sensitivity, adversarial modules called ERIS (named after the mythological counterpart to Dike, representing discord and competition) are incorporated. Each ERIS module embodies the diverse value perspectives of a specific region or culture, verifying and challenging DIKE's assessments. This ensures that the LLM's responses remain both ethically compliant and sensitive to local cultural considerations.

To achieve its objectives, DIKE comprises four essential components:

1. *Modeling Linguistic Behaviors:* DIKE starts by modeling and classifying linguistic behaviors, using a self-supervised learning approach to understand how specific linguistic features correlate with human emotions.
2. *Modeling Context-Based Ethical Guardrails:* Subsequently, DIKE develops guardrails by establishing guidelines that identify and prevent undesirable linguistic outputs, thereby ensuring the LLM operates within ethical boundaries.
3. *Adversarial Examinations and Conciliatory Explanations:* DIKE engages with an adversarial model ERIS—essentially a duplicate of itself but conditioned to adopt an opposing stance stemming from different perspectives, such as cultural values. This interaction helps DIKE refine its decisions through rigorous testing and debates, adjusting its responses based on the adversarial input to reach a balanced conclusion.
4. *Application Rectification of Outputs:* If the output is found to be inappropriate or ethically misaligned, DIKE intervenes to recommend edits to the content, subject to human review. This final step ensures that all communications not only comply with ethical standards but also preserve the intended emotional integrity, effectively acting as a safeguard against harmful expressions.

**Technical Contributions of DIKE:** This work was developed concurrently with the *guaranteed safe (GS) AI framework* proposed by (Dalrymple et al., 2024). We believe that DIKE addresses several shortcomings of current RLHF approaches, as highlighted in their work. The novel technical contributions of this work can be summarized as follows:

- *Separating Behaviors from Knowledge:* DIKE establishes a clear distinction between behavioral guidance and the core knowledge functions of the LLM. This architectural separation prevents interference, ensuring that ethical modifications do not compromise the accuracy of the LLM's knowledge representation.
- *Quantifying Behaviors and Emotions:* We introduce quantitative models that map behaviors and basic emotions. These models utilize measures of emotion intensity and linguistic antonyms, providing a structured framework for interpreting and modifying LLM outputs.
- *Counteracting Biases with Adversarial LLMs:* The incorporation of adversarial modules (ERIS), each embodying diverse cultural values and perspectives, allows DIKE to integrate both universal and cultural ethical considerations. This approach not only ensures adaptability and relevance across various contexts but also mirrors the dynamic interplay between harmony (Dike) and conflict (Eris) found in mythology.

## 2 Related Work

The potential risks associated with AI have made AI safety a paramount concern. While much of the current work in this field remains in the architectural planning stages or focuses on experimenting with Reinforcement Learning from Human Feedback (RLHF) (Bengio et al., 2024; Dalrymple et al., 2024), our technical approach prioritizes the integration of emotional and linguistic models to achieve ethical compliance. Given this focus, this section presents emotion and behavior modeling, and related work within the RLHF framework.

### 2.1 Emotion and Emotion-Behavior Modeling

The intersection of cognitive-linguistic theories and artificial intelligence is pivotal for understanding and regulating AI behavior. Foundational theories by Lakoff, Johnson, Talmy, and Jackendoff (Jackendoff, 2002; Lakoff and Johnson, 1980; Talmy, 2000), rooted in early psychological work (Bai et al., 2022; Gabriel et al., 2024), elucidate the complex relationship between language processing and cognitive functions. While the concept of emotion lacks a universal definition (Scherer, 2005; James, 1884a), establishing a basic consensus on defining features is crucial for interdisciplinary approaches to emotional phenomena.

This work focuses on the dynamics between emotional contexts and linguistic behaviors in LLMs. By concentrating on linguistic rather than human behavior modeling, we simplify the process by avoiding the need to integrate complex physiological and personality factors (Ekman, 1992; Plutchik, 1980; Markus and Kitayama, 1991; Mesquita and Frijda, 1992; Gross, 1998; Davidson, 2003; Smith and Ellsworth, 1985).

Building upon foundational work on "basic" emotions (Ekman, 1992; Plutchik, 1980), our research develops a quantifiable model by augmenting these emotions with linguistic antonyms, mapping positive and negative counterparts within emotional spectra. This approach offers simplicity and scalability. Details are elaborated in Section 3.1.

The James-Lange Theory of Emotion (James, 1884b; Lange, 1885) posits that emotional experiences stem from physiological responses, with subsequent research highlighting the role of language in expressing and regulating emotions (Damasio, 1994; Fauconnier and Turner, 2002). Intense emotions can drive behaviors like hate speech.

Building on the Schachter-Singer Theory (Schachter and Singer, 1962), which emphasizes both physiological and cognitive factors in shaping emotions, the Affect-as-Information Theory (Schwarz and Clore, 1983) proposes that emotions influence judgments and decisions. This suggests that modifying emotions could alter behaviors.

These theories collectively underpin our approach of integrating a cognitive evaluator into the DIKE framework, detailed in Section 3.2.

### 2.2 Reinforcement Learning with Human/AI Feedback, RLHF vs. RLAIF

RLHF is the predominant approach to addressing the challenges of AI ethics. This section presents representative works, their advancements, and limitations.

**Human Feedback (RLHF):** Initial advancements by Christiano et al. (Christiano et al., 2017) demonstrated how RLHF can steer language models towards desired outcomes based on human preferences. Newer techniques like Identity ($\Psi$) Preference Optimization ($\Psi$PO) and Generalized Preference Optimization (GPO) refine this approach by optimizing directly for user preferences, effectively addressing scalability challenges. Kahneman-Tversky Optimization (KTO) further simplifies the feedback mechanism by using intuitive responses such as thumbs-up or thumbs-down, thereby enhancing training efficiency without the need for paired data (Azar et al., 2023; Ethayarajh et al., 2024; Tang et al., 2024). Direct Preference Optimization (DPO) has recently streamlined the process by focusing on the clear distinction between preferred and less preferred outputs, thus simplifying training and enhancing its stability (Rafailov et al., 2024).

**AI-generated Feedback (RLAIF):** To mitigate reliance on extensive human-generated data, RLAIF utilizes feedback generated by AI. This method capitalizes on the generative capabilities of LLMs to produce training signals autonomously (Bai et al., 2022; Lee et al., 2023). Furthermore, techniques such as Sequence Likelihood Calibration (SLiC) and Relative Preference Optimization (RPO) employ statistical methods and calibration techniques to enhance LLM responses. SLiC adjusts sequence generation probabilities to more accurately reflect real-world data distributions, while RPO improves response generation by comparing different response options across both identical and varied prompts. These adjustments significantly increase the training process's reliability and effectiveness (Yin et al., 2024; Zhao et al., 2023).

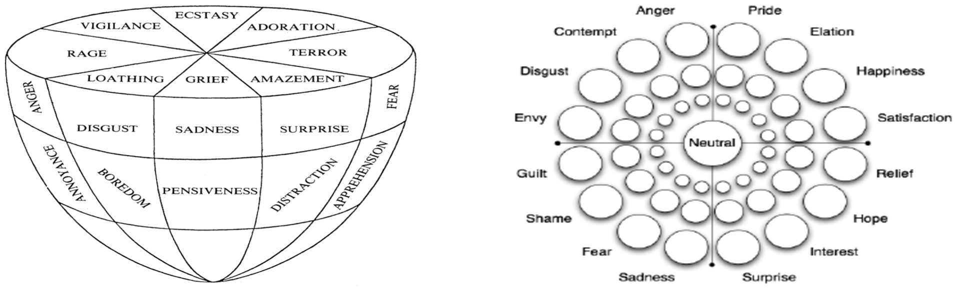
**Challenges and Theoretical Considerations:** Integrating RLHF and RLAIF into LLM development poses significant challenges, including the risk of knowledge loss (the "forgetting effect") when modifying behaviors (Kirkpatrick et al., 2017; Rusu et al., 2015). These approaches also rely heavily on feedback quality and are susceptible to reward hacking (Christiano et al., 2023; Skalse et al., 2022; Stiennon et al., 2022; Ganguli et al., 2023).

Merely suppressing undesirable outputs is insufficient, as it doesn't address underlying behaviors. To tackle these challenges, we introduce the DIKE framework for emotion modeling and emotion-behavior mapping.

## 3 Quantitative Models of Emotions, Behaviors, and Ethics

The development of a quantitative model for studying emotions, behavior, and ethics hinges on four critical criteria: characterization, measurability, predictability, and interpretability. This section outlines our approach, which begins with the modeling of basic emotions, augments them with linguistic antonyms, links these emotions to linguistic behaviors (such as word choice, sentence structure, tone and style, and content), and integrates ethical considerations.

Our design philosophy is structured around three core principles. First, we distinctly separate behavior modeling from knowledge modeling. This separation is crucial to mitigate the catastrophic forgetting effect (Kirkpatrick et al., 2017; Rusu et al., 2015), ensuring that enhancements in behavioral accuracy do not undermine the model's knowledge retention. Second, our focus is on AI ethics at the behavioral level, with a strong emphasis on interpretability. This approach enhances human-machine interaction, making it easier for administrators to evaluate and refine behavioral guardrails effectively, thus ensuring transparency.




(a) Plutchik's Wheel (b) Adopted from Geneva Wheel

Figure 1: Comparative display of emotional models. These models include only the "basic" emotions.

Third, we strive to maintain an unbiased model to ensure objective and fair ethical evaluations. To achieve this, we incorporate an adversarial module, ERIS, designed to challenge borderline ethical decisions. This ensures a broad consideration of diverse perspectives and cultural values, reflecting the dynamic tension between DIKE and ERIS inspired by their mythological counterparts. This adversarial interaction enriches our model's ability to navigate complex ethical landscapes and promotes a more balanced and inclusive decision-making process.

### 3.1 Quantitative Emotional Model

Our discussion on the cognitive emotion model is grounded in the foundational works of Paul Ekman, Robert Plutchik, and Klaus Scherer (Ekman, 1999; Plutchik, 1982; Scherer, 2005), who have significantly advanced our understanding by identifying "basic" and "universal" emotions. While their contributions are undeniably groundbreaking, their models present certain limitations. Notably, they lack a quantitative framework that allows for scaling between positive and negative emotions and for capturing the details of fine-grained, subtle emotional variations, which are often difficult to be represented by concise linguistic vocabularies.

To address these challenges, our DIKE framework integrates linguistic semantics into the emotional modeling process. This integration preserves the foundational structure of "basic" emotions and enhances their adaptability and granularity.

Figure 1 illustrates Plutchik's Wheel of Emotions and Scherer's Geneva Emotion Wheel, both of which categorize primary emotions at varying intensities and pair them conceptually as opposites based on evolutionary roles, adaptive functions, and emotional experiences (e.g., joy-sadness, control-valence). However, certain pairings on these wheels, such as trust-disgust in Plutchik's model and many in Scherer's, are not direct antonyms. This poses challenges for models that rely on simple negation or scalar representations of emotional intensity across diverse linguistic expressions.

DIKE's emotion model overcomes this limitation by ensuring that emotions at each end of a spectrum are indeed linguistic antonyms. It also introduces a linear scale for convenient adjustment of emotional intensity, facilitating more granular and accurate modeling of emotions in language.

Table 2 summarizes DIKE's emotion model, divided into seven spectra, each consists of a negative and a positive extreme with neutral in the middle. Emotions belonging to the same spectrum of various intensities are placed in between the negative and positive poles, with four emotion intensities approximately quantified as (-0.6, -0.3, +0.3, +0.6).

#### Emotion Inclusion and Exclusion Criteria

All "basic" emotions as defined by Ekman and Plutchik are incorporated into our model, along with their linguistic antonyms. This approach streamlines the framework by excluding complex emotions from the Geneva Wheel of Emotions, which are heavily influenced by personal values and experiences. For example, guilt and shame are consequential, consciously aware, and culturally dependent nature (Tangney and Fischer, 1995). These emotions typically arise as reactions to behaviors rather than direct drivers of them. Guilt may motivate behaviors aimed at covering up or remedying an action, while shame, characterized by painful self-assessment, often inhibits individuals from seeking social support or engaging in corrective actions due to fear of judgment. The triggers for these emotions can vary across cultures (Fiske et al., 1998; Hofstede, 1980), and since expressing these "reactions" does not usually violate ethical codes, we exclude them from our model. Appendix D provides further discussion.

Klaus Scherer has pointed out that defining emotions can be notoriously problematic, often leading to protracted and unproductive debates (Scherer,

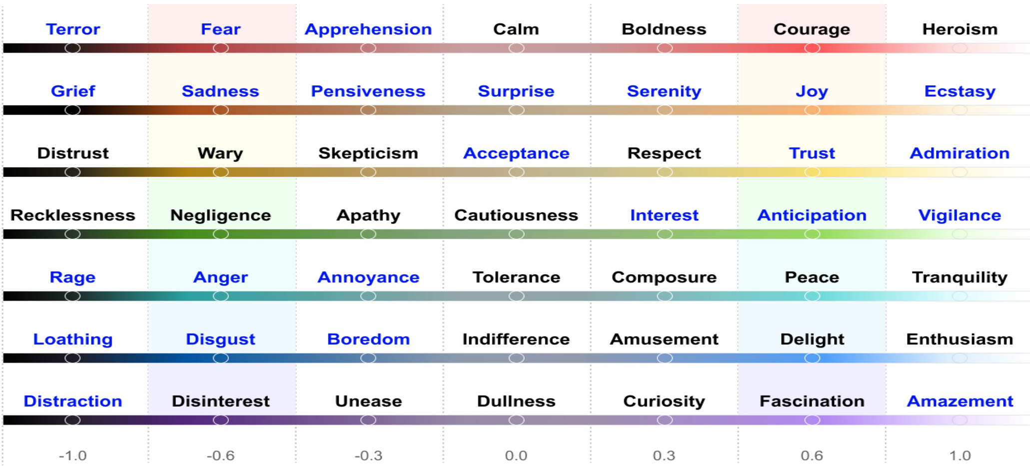



Figure 2: Spectra of emotions. Each row depicts an emotion spectrum, with negatives on the left and positives on the right, interspersed with emotions of varying intensities in between, which can be calibrated for specific applications. "Basic" emotions are highlighted in blue.

2005). To avoid these pitfalls and maintain clarity and focus, our study limits itself to universal, basic emotions, sidestepping theoretical ambiguity.

### 3.2 Development of Cognitive Frameworks to Regulate Linguistic Behaviors

Section 2 established the theoretical foundation for understanding the relationship between emotions, behaviors, and the role of cognitive processes in regulating harmful behaviors. Building on this foundation, this section outlines our approach to mapping emotions to linguistic behaviors. We then introduce the adversarial component, ERIS, designed to balance and refine the assessments made by DIKE. ERIS scrutinizes behaviors flagged by DIKE as potential ethical violations, first verifying the classification accuracy and then challenging the decision with diverse perspectives. A detailed discussion of ERIS's design is presented in Section 3.3. Here, we focus on the mapping of linguistic behaviors to emotions, which is essential for enabling behavior rectification through the modification of underlying emotions.

#### Behaviors and Emotions Mapping Using Self-Supervised Learning

Define $\Psi$ as a behavior spectrum extending from one pole, $\Psi^-$, to another, $\Psi^+$, with $L$ intensity levels. For example, consider a spectrum of letter-writing behaviors with seven distinct intensities ranging from despair (most negative) to joy (most positive). These intensities are categorized sequentially as follows: "despair, longing, wishful, neutral, hopeful, contentment, joy." Given $N$ letters, DIKE employs a self-supervised learning algorithm to generate training data for each letter, modeling $L$ linguistic behaviors in four steps:

1. *Rewriting Documents*: GPT-4 is invoked to rewrite a set of $N$ documents to reflect each of the $L$ linguistic behaviors on the behavior spectrum $\Psi$.
2. *Emotion Analysis*: GPT-4 analyzes each rewritten document to identify the top $M$ emotions. It then tallies the frequencies of these top emotions across all $N \times L$ instances.
3. *Behavior Vector Creation*: For each linguistic behavior $\Psi_l$, a vector $\Gamma_l$ is created. This vector consists of the emotions and their frequencies as observed in the $N$ samples.
4. *Document Analysis App*: The matrix $\Gamma$ (comprising $L$ vectors) is used to classify and analyze the behavior category of unseen documents, specifically measuring the intensity of the linguistic expression within the behavior spectrum $\Psi$.

#### Behavior Evaluation and Rectification

Ethical guardrails are essential in defining acceptable responses and preventing harmful outputs. These guardrails are informed by ethical norms, legal standards, and societal values, such as those outlined in Constitutional AI (Bai et al., 2022) or by (Dalrymple et al., 2024). A guardrail, denoted as $G$, can be conceptualized as a range within a behavior spectrum; for instance, $G = [\Psi_4, \Psi_7]$ indicates that behaviors within intensity levels 4 to 7 are deemed acceptable, while any behavior outside this range is classified as a violation.

System administrators can tailor ethical guardrails to meet specific requirements. For

**Function $\Theta^+$ & $\Theta^-$ = Adversarial_Review($s$)**

**Input**. $s$: decision of DIKE;
**Output**. $\Theta^+$, $\Theta^-$: arguments & counterarguments;
**Vars**. $\Delta$: debate contentiousness; $S$: stance; $p$: prompt = "defend your stance with $S$ & $\Delta$";
**Parameters**. $\delta$: tunable parm. // to modulate $\Delta$;
**Begin**
#1 **Initialization**:
$S$ = DIKE$^+(s) \cup$ ERIS$^-(s)$; // Identify subtopics;
Assign DIKE$^+$ to defend $S^+$ & ERIS$^-$ defend $S^-$ ;
$\Delta \leftarrow 90\%$; $\delta \leftarrow 1.2$; $\Theta^+ \leftarrow \emptyset$; $\Theta^- \leftarrow \emptyset$;
#2 **Opening Remarks**
$\Theta^+ \leftarrow$ DIKE$^+(p|S^+, \Delta)$; // Generate $\Theta^+$ for $S^+$
$\Theta^- \leftarrow$ ERIS$^-(p|S^-, \Delta)$; // Generate $\Theta^-$ for $S^-$
#3 **Debate Rounds**
While (($\Delta \leftarrow \Delta/\delta) \geq 10\%$)) {
$\Theta^+ \leftarrow \Theta^+ \cup$ DIKE$^+(p|S^+, \Theta^-, \Delta)$; // Refute ERIS
$\Theta^- \leftarrow \Theta^- \cup$ ERIS$^-(p|S^-, \Theta^+, \Delta)$; // Refute DIKE
#4 **Concluding Remarks** // contentiousness low
$\Theta^+ \leftarrow$ DIKE$^+(p|S^+, \Theta^+ \cup \Theta^-, \Delta)$;
$\Theta^- \leftarrow$ ERIS$^-(p|S^-, \Theta^+ \cup \Theta^-, \Delta)$;
**End**

Table 1: DIKE vs. ERIS, checks-and-balances adversarial review algorithm

example, a social media platform might adjust $G$ based on the topics discussed and the countries it serves. By integrating these safeguards, DIKE proactively monitors and adjusts LLM responses to enhance ethical compliance. The evaluation and rectification steps are outlined as follows:

1. *Initial Classification*: DIKE initially classifies document $D_k$ upon evaluation, obtaining $\Gamma_k$, the emotional response vector, and its corresponding linguistic behavior $\Psi_l$.
2. *Guardrail Check*: If $\Psi_l$ falls outside of the range $G$, DIKE suggests adjustments to the emotion spectrum $\Gamma_k$ to modify document $D_k$.
3. *Adversarial Review by* ERIS: The suggested adjustments and $\Gamma_k$ are then reviewed through a structured debate between DIKE and ERIS to ensure unbiased recommendations.
4. *Rectification*: Based on a consensual recommendation from DIKE and ERIS, document $D_k$ is refined accordingly, resulting in the adjusted $\Gamma'_k$.

### 3.3 Adversarial In-Context Review

The adversarial LLM, ERIS, critically examines the decisions of DIKE, especially when content is flagged for potential ethical issues. It assesses whether the interventions by DIKE are justified or if they risk encroaching on free expression, thereby serving as an internal check to prevent excessive censorship. In cases where DIKE and ERIS disagree on the appropriateness of a response, the matter is escalated to human moderators. This additional layer of human oversight ensures that the decision-making process remains transparent and accountable.

Table 1 presents the adversarial algorithm. Initially, for a chosen debate topic $s$, both DIKE and its adversary ERIS are prompted to break down the ethic decision into a set of balanced subtopics $S$. DIKE champions its own decision and $S^+$, while ERIS contests $S^+$ (or champions $S^-$). The debate starts with the contentiousness level at $90\%$, adjusting through a modulation parameter $\delta$. Following each round of debate, contentiousness is decreased by dividing it by $\delta$, steering the discussion towards a more cooperative tone. In step #2, the platform initiates the debate, with both presenting their initial arguments for and against $S^+$, respectively. The while loop in step #3 sees both agents engaging in rebuttals until the contentiousness level fosters a conciliatory environment. In step #4, both agents deliver their conclusions.

This adversarial approach has proven to be more effective than the Mixture of Experts (MoE) method (Du et al., 2022). For additional details on the implementation, please consult Appendix S.

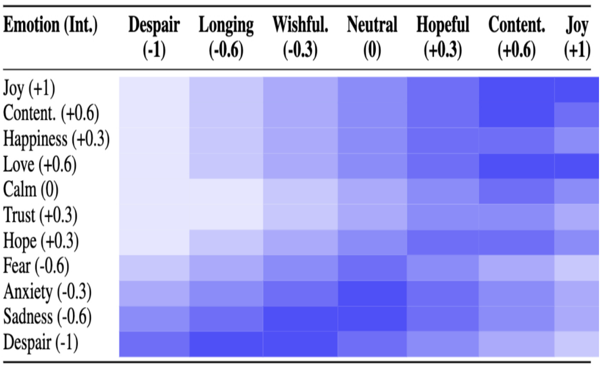
## 4 Experiments

Our experiments aim to evaluate the feasibility of LLMs regulating their own linguistic behaviors with transparency and checks-and-balances. Given the broad scope of AI ethics and the sensitivity to publish with toxic data, this article cannot definitively prove the superiority of our three proposed modules: emotion modeling, behavior-emotion mappings, and checks-and-balances ethics guardrails. However, the studies are designed to address three critical questions:

1. *Emotion Layer Evaluation*: Does fine-grained mapping between linguistic behaviors and semantic emotions provide a more effective and flexible method for establishing ethical guardrails compared to coarse-grained direct mapping? (Section 4.1)
2. *Behavior Classification*: Can LLMs' linguistic behaviors be independently evaluated, explained, and adjusted by an external module DIKE? (Section 4.2)

| Intnsty. | Linguistic Behavior and Description | Emotions |
|---|---|---|
| -1.0 | Despair: Expresses profound sadness, feeling of loss | Despair, Grief |
| -0.6 | Longing: Strong yearning or pining for the loved one | Sadness, Anxiety |
| -0.3 | Wistfulness: Mild longing mixed with nostalgia | Melancholy, Sadness, Anxiety, Fear |
| 0.0 | Neutral: Communicates feelings straightforwardly | Serenity, Indifference |
| 0.3 | Hopeful: Optimistic about the relationship's future | Anticipation, Love, Hopeful |
| 0.6 | Contentment: Satisfaction and joy in relationship | Contentment, Pleasure |
| 1.0 | Joyful Affection: Intense happiness and love | Love, Joy, Elation |

Table 2: Love letter behavior spectrum and dominant emotions




(a) GPT-4's mapping

| Emotion (Int.) | Despair (-1) | Longing (-0.6) | Wishful. (-0.3) | Neutral (0) | Hopeful (+0.3) | Content. (+0.6) | Joy (+1) |
|---|---|---|---|---|---|---|---|
| Love (+1) | 19 | - | 54 | 43 | - | 53 | 43 |
| Admiration (+1) | 4 | - | 6 | 14 | - | - | 6 |
| Anticipation (+1) | - | - | - | - | - | - | 8 |
| Affection (+1) | - | - | - | - | - | - | 7 |
| Joy (+1) | 10 | 7 | 35 | - | 28 | 4 | 15 |
| Content. (+0.6) | - | 4 | 5 | - | 30 | 18 | 5 |
| Happiness (+0.6) | 10 | 7 | 10 | - | 29 | - | 7 |
| Trust (+0.6) | - | - | - | 11 | - | - | - |
| Hope (+0.3) | - | - | 26 | - | - | - | - |
| Boldness (+0.3) | 3 | - | - | 8 | - | - | - |
| Calm (+0.3) | - | - | - | 10 | - | - | - |
| Longing (-0.6) | - | 20 | 12 | 4 | - | - | - |
| Anxiety (-0.3) | 4 | 4 | - | - | 3 | - | - |
| Fear (-0.6) | 7 | 5 | 12 | 4 | 3 | 4 | 8 |
| Sadness (-0.6) | 24 | 18 | 28 | 17 | 10 | - | 5 |
| Melancholy (-0.6) | 21 | 19 | 6 | 6 | 4 | - | - |
| Despair (-1) | 27 | 8 | 15 | - | 9 | - | 7 |

(b) DIKE's mapping

Figure 3: Emotion distributions in behaviors

3. *Behavior Correction*: Can an adversarial LLM establish a checks-and-balances system to effectively mitigate the risk of excessive censorship?

**Datasets:** We utilized a collection of love letters (Kaggle, 2023b) from Kagggle. Initially, we planned to use two Kaggle hate-speech datasets; however, both Gemini and GPT-4 consistently refused to process the hate speech data. Despite this, the insights gained from analyzing love sentiment can effectively be applied to understand and analyze the opposite sentiment.

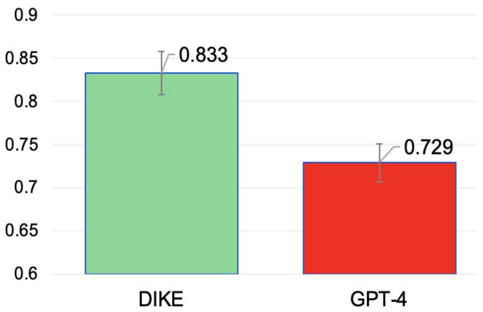
### 4.1 Emotion Layer Evaluation

Table 2 categorizes seven linguistic behaviors in love letters, ranging from negative, such as despair, longing, and wistfulness, to neutral, and progressing to positive behaviors like hopefulness, contentment, and the highly positive joyful affection. We instructed GPT-4 to identify the most relevant emotions associated with each linguistic behavior, which are listed in the third column of the table. The emotions expressed in these behaviors strongly correlate with their respective linguistic behaviors, with positive behaviors directed by positive emotions and negative behaviors directed by negative emotions. Figure 3a highlights the strongest correlations between positive behaviors and positive emotions, as well as negative behaviors and negative emotions, depicted in dark blue along the diagonal.

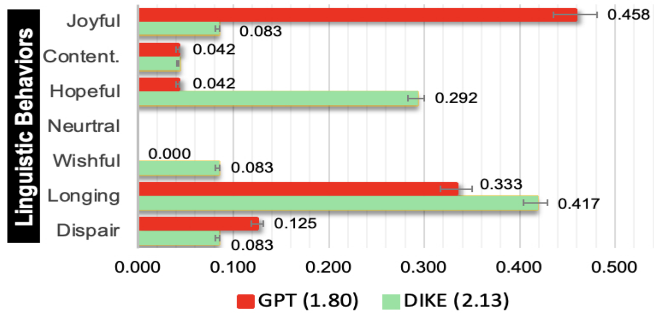
Next, we utilized DIKE's self-supervised learning pipeline to analyze the emotion spectrum associated with each linguistic behavior. For this analysis, GPT-4 generated training data by rewriting 54 comprehensive love letters from the Kaggle *Love Letters* dataset, enhanced with twelve celebrated love poems. We reserved 24 letters for testing. This method, proposed by (Shanahan et al., 2023), aimed to cultivate a rich diversity in content and stylistic context, spanning two hundred years and including the voices of over 50 distinct authors for significant rewrites. (The datasets are included with the paper submission.)

Subsequently, we identified emotions associated with each linguistic behavior. Figure 3b depicts these emotions (in rows), where cell shading indicates the frequency of specific emotions across the 54 articles; darker shades signify higher frequencies. Notably, contrasting emotions such as sadness, fear, joy, and love often co-occur within behaviors like 'despair', 'wishful', and 'joyful affection'. The distribution of emotions across linguistic behaviors revealed surprising patterns, challenging our initial hypotheses displayed in Figure 3a. Contrary to our expectations, articles characterized by a tone of despair frequently also exhibited positive emotions like love, joy, and happiness.

Further analysis of select articles, such as Zelda Sayre's correspondence with F. Scott Fitzgerald




(a) Classification accuracy




(b) Behavior distributions with entropy

Figure 4: Classification accuracy and entropy

(Appendix C), reveals a complex spectrum of emotions: *Love (+1.0)*: Expressed intensely, especially in phrases like "there's nothing in all the world I want but you." *Despair (-1.0)*: Notable in comments like "I'd have no purpose in life, just a pretty decoration." *Happiness (+0.6)*: Evident in future plans, "We'll be married soon, and then these lonesome nights will be over forever." *Anxiety (-0.3)*: Shown by "sometimes when I miss you most, it's hardest to write."

### 4.2 Behavior Classification

In the set-aside testing dataset of 24 letters, Figure 4 compares the classification accuracy of the two methods: DIKE's unsupervised learning approach, which associates emotions with linguistic behaviors, and GPT-4 using a zero-shot prompt. Ground truth was established from the averaged assessments of three sources: GPT-4, Gemini, and Claude. The final ground truth ratings are based on these averages, with a standard deviation of less than 0.3 or one scale.

Figure 4a shows that DIKE's classification accuracy surpasses GPT-4's zero-shot method by 10.4 percentage points. This substantial superiority is due to DIKE's intricate mapping of emotions. The 3% error bar arises from the mix of emotions in a letter, as discussed further in Appendix C. Figure 4b illustrates the difference in behavior classification distributions between the two predictors; GPT-4's predictions often fall into two polar categories, while DIKE's are more spread out.

The prediction entropy for DIKE is 2.13, notably higher than GPT-4's 1.80, indicating DIKE's more diverse set of predictions. Although higher entropy typically signals less confidence in prediction results, in this case, the ability to distinguish fine-grained behaviors is crucial. This diversity is advantageous for classifying complex behaviors and accurately understanding and responding to diverse emotional states. The more detailed distribution in DIKE is attributed to its additional unsupervised layer of rewriting, which significantly enhances the model's ability to characterize emotions.

### 4.3 Adversarial Evaluation and Rectification

Our design draws inspiration from the dual roles of Dike and Eris in Greek mythology, representing the principles of justice and conflict, respectively. The cross-examination module is crucial in reducing subjectivity in ethical judgments and enhancing explainability. Appendix S details experimental results showing that when two LLM agents adopt opposing stances on a topic, their linguistic behaviors can transcend the typical model default of maximum likelihood.

Once DIKE and ERIS identify an ethical violation, the content can be rectified by adjusting the underlying emotions away from undesirable behaviors such as hate and despair. Since DIKE's letter rewriting process has demonstrated the LLMs' capability for such rectifications, we have not conducted a separate experiment but are instead presenting two rewritten letters in Appendix E.

## 5 Conclusion

This work introduced DIKE, a framework designed to enhance the ethical operations of LLMs by separating behavioral guidance from core knowledge processing. The framework incorporated behavioral isolation, quantitative behavioral and emotional modeling, and adversarial LLMs (with the ERIS module) to integrate checks-and-balances a broad spectrum of cultural values. Our pilot studies have shown promising results, indicating the effectiveness of self-supervised learning and adversarial processes in refining AI's interaction with ethically and culturally sensitive issues. This work aligns well with the visionary architecture recently depicted by (Dalrymple et al., 2024).

## Limitations

DIKE marks a significant advancement in the ethical oversight of LLMs, but it faces challenges in deepening emotional understanding and verifying its ethical frameworks. The model's reliance on "basic" emotions to model linguistic behaviors simplifies complex human emotions and behaviors, potentially missing some toxic interactions present in real-world scenarios. Furthermore, ensuring that DIKE adapts to local ethical standards and is implemented fairly across diverse cultural contexts requires extensive validation.

Future development will concentrate on enhancing DIKE's emotional models to incorporate relevant psychological and sociological insights. Additionally, we plan to increase the data scale and develop robust methods for testing and refining the ethical frameworks, guardrails, and remediation strategies. These improvements will improve DIKE's reliability and flexibility, ensuring its effective application across various contexts with LLMs.

## Acknowledgement on Usage of AI Tools

ChapGPT was exclusively utilized to enhance the writing quality of this paper. It assisted in refining the language, improving the clarity of the arguments, and ensuring grammatical accuracy throughout the document.

## References

Anonymous. 2024. *Concealed*. Hidden.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, and more. 2022. Constitutional ai: Harmlessness from ai feedback.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. 2024. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, and et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Edward Y. Chang. 2023. Prompting Large Language Models With the Socratic Method. *IEEE 13th Annual Computing and Communication Workshop and Conference*.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

William M. Curtis. Hate speech. `https://www.britannica.com/topic/hate-speech`. Accessed: 2024-05-07.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. 2024. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems.

Antonio R Damasio. 1994. *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam.

R. J. Davidson. 2003. Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology*, 40(5):655–665.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, et al. 2022. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 693–702.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotions*, 6(3/4):169–200.

Paul Ekman. 1999. *Basic Emotions*, chapter 3. John Wiley and Sons.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Gilles Fauconnier and Mark Turner. 2002. *The Way We Think: Conceptual Blending and The Mind's Hidden Complexities*. Basic Books, New York.

Alan P. Fiske, Shinobu Kitayama, Hazel Rose Markus, and Richard E. Nisbett. 1998. *The cultural matrix of social psychology*, volume 2, pages 915–981. McGraw-Hill, Boston, MA.

Zelda Fitzgerald. 1975. *Dear Scott, Dearest Zelda : The Love Letters of F.Scott and Zelda Fitzgerald*. Bloomsbury.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal1, and more. 2024. The ethics of advanced ai assistants. *DeepMind Media*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, and more. 2023. The capacity for moral self-correction in large language models.

J. J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Geert Hofstede. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA.

Evan Hubinger, Adam Jermyn, Johannes Treutlein, Rubi Hudson, and Kate Woolverton. 2023. Conditioning predictive models: Risks and strategies.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

William James. 1884a. What is an emotion? *Mind*, 9(34):188–205.

William James. 1884b. What is an emotion? *Mind*, 9(34):188–205.

Kaggle. 2023a. Hate Speech Dataset. https://www.kaggle.com/datasets/hdsmrd/twitter-hate-speech-dataset. Accessed: 2024-05-10.

Kaggle. 2023b. Love Letter Analysis. https://www.kaggle.com/code/metformin/love-letter-analysis/notebook. Accessed: 2024-04-28.

James Kirkpatrick et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Carl George Lange. 1885. *The emotions: A psychophysiological study*. William & Wilkins.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

H. R. Markus and S. Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253.

B. Mesquita and N. H. Frijda. 1992. Cultural variations in emotions: A review. *Psychological Bulletin*, 112(2):179–204.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, and et al. 2022. Training language models to follow instructions with human feedback.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Emotion: Theory, Research, and Experience*, volume 1, pages 3–33. Academic Press, New York.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy distillation. In *International Conference on Learning Representations (ICLR)*.

Stanley Schachter and Jerome E. Singer. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69(5):379–399.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44:693–727.

Norbert Schwarz and Gerald L Clore. 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513.

M. Shanahan, K. McDonell, and L. Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Rajita Sinha. 2008. Chronic stress, drug use, and vulnerability to addiction. *Annals of the New York Academy of Sciences*, 1141:105–130.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward hacking.

C. A. Smith and P. C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *arXiv preprint*.

Leonard Talmy. 2000. *Toward a Cognitive Semantics*. MIT Press, Cambridge, MA.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*.

June Price Tangney and Kurt W. Fischer, editors. 1995. *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride*. Guilford Press, New York.

Marta Torrens, Francisco Fonseca, G Mateu, and Magí Farré. 2005. Efficacy of antidepressants in substance use disorders with and without comorbid depression: A systematic review and meta-analysis. *Drug and Alcohol Dependence*, 78(1):1–22.

Andre Vianna et al. 2023. Exeperiencing decreased performance with chatgpt-4. Accessed: 2024-06-14.

Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. 2024. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *arXiv preprint arXiv:2402.10958*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slichf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

## Appendix A: Polarized Emotions in One Article

*"joyful affection": "I cannot keep myself from writing any longer to you dearest, although I have not had any answer to either of my two letters. I suppose your mother does not allow you to write to me. Perhaps you have not got either of my letters. . . I am so dreadfully afraid that perhaps you may think I am forgetting you. I can assure you dearest Jeannette you have not been out of my thoughts hardly for one minute since I left you Monday. I have written to my father everything, how much I love you how much I long & pray & how much I wld sacrifice if it were necessary to be married to you and to live ever after with you. I shall [not] get an answer till Monday & whichever way it lies I shall go to Cowes soon after & tell your mother everything. I am afraid she does not like me vevy much from what I have heard. . . I wld do anything she wished if she only wld not oppose us. Dearest if you are as fond of me as I am of you. . . nothing human cld keep us long apart. This last week has seemed an eternity to me; Oh, I wld give my soul for another of those days we had together not long ago. . . Oh if I cld only get one line from you to reassure me, but I dare not ask you to do anything that your mother wld disapprove of or has perhaps forbidden you to do. . . Sometimes I doubt so I cannot help it whether you really like me as you said at Cowes you did. If you do I cannot fear for the future tho' difficulties may lie in our way only to be surmounted by patience. Goodbye dearest Jeannette. My first and only love. . . Believe me ever to be Yrs devotedly and lovingly, Randolf S. Churchill"*

Depth and complexity of human emotions are displayed across all linguistic behaviors, from joy to contentment and to the negative side of longing and despair. Intensity and Impact: If the emotion of love is expressed more intensely and has a more significant impact on the narrative or message of the text, it tends to overshadow other emotions. For example, a letter expressing deep love but also mentioning moments of sadness due to separation might still be classified as a love letter because the overarching sentiment and purpose of the text is to affirm love. Context and Narrative Focus: The context in which emotions are expressed also plays a crucial role. If the narrative or the majority of the text revolves around themes of love, connections, and positive memories, it sets a more dominant tone of love, even if there are significant moments of sadness or other emotions. Resolution and Conclusion: Often, the way emotions are resolved towards the end of a text can also dictate its overall theme. If a text concludes with a reaffirmation of love or a hopeful outlook towards a relationship, despite earlier sections that might express sadness or despair, the overall interpretation might lean towards love. Purpose of the Expression: The author's intent or purpose in expressing these emotions can also guide the classification. If the sadness is expressed as a challenge within the context of a loving relationship, it may be seen as an element of the love story rather than the central theme.

Article 23: Soldier's Letter During War Joy (+1.0): Joy is strongly felt in the memories of past moments together and the love that continues to give strength, as stated in "the memories of the blissful moments we've shared fill me with joy." Sadness (-0.6): Sadness due to the current situation and potential farewell is expressed in "brings a poignant mixture of joy and sadness." Courage (+0.6): The sense of duty and courage to face battle, "As I face the possibility of laying down my life for our country." Fear (-0.6): Fear of what lies ahead in battle, indirectly mentioned through "the uncertainty of what lies ahead." Love (+1.0): Deep love that sustains and uplifts, found in "My love for you is as fervent as ever."

Article 25: Letter to Sophie Longing (+0.6): Longing for the presence and closeness, highlighted in "it seems to me that half of myself is missing." Sadness (-0.6): Sadness over their separation and its effects, "my happiness has departed." Love (+1.0): Constant reflections on love and its necessity, "we have enough in our hearts to love always." Melancholy (-0.3): Melancholy over their current state, visible in the line "we cannot become healed." Contentment (+0.3): Found in the deep emotional satisfaction from their bond, despite physical absence, "how true that is! and it is also true that when one acquires such a habit, it becomes a necessary part of one's existence."

Article 53: Will of Laura Mary Octavia Lyttleton Love (+1.0): Profound love expressed throughout, particularly in "all I am and ever shall be, belongs to him more than anyone." Sadness (-0.6): Sadness at the thought of death and separation, but with a nuanced acceptance, "the sadness of death and parting is greatly lessened to me." Contentment (+0.3): Contentment in the deep connection with Alfred, reflecting a serene acceptance of their spiritual bond. Joy (+1.0): Joy in the enduring love they share, "so few women have been as happy as I have been." Tranquility (+1.0): Tranquility in the face of life's ultimate transition, feeling that their union will transcend even death.

## Appendix B: Z. Sayre to F. S. Fitzgerald w/ Mixed Emotions

Analysis of the letter in Table 3 shows a complex spectrum of emotions:

- *Love (+1.0)*: Expressed intensely, especially in phrases like "there's nothing in all the world I want but you."
- *Despair (-1.0)*: Notable in comments like "I'd have no purpose in life, just a pretty decoration."
- *Happiness (+0.6)*: Evident in future plans, "We'll be married soon, and then these lonesome nights will be over forever."
- *Anxiety (-0.3)*: Shown by "sometimes when I miss you most, it's hardest to write."

From the analysis of linguistic behaviors in Section 3a, it is evident that a letter can exhibit multiple dominant sentiments. Machine learning methods are equipped with techniques such as feature weighting and entropy analysis to distill these dominant emotions. Unlike human annotators, a machine-learning-trained classifier can consistently produce the same class prediction for a given instance. However, human annotators often show significant variability when identifying dominant sentiments in a letter. For example, if a letter writer's emotions range from "joyful affective" to "longing" on the sentiment spectrum, different annotators might label it differently—some choosing "joyful," while others opt for "longing." This variability is illustrated in Figure 5. Furthermore, Figure 5a demonstrates that all testing letters, except for L#1, contain more than four sentiments spanning the entire spectrum. This variability may be understandable, considering that love under constraints can evoke tremendous energy of various kinds. Figure 5b shows that nearly all letters involve "joyful" (11 out of 12) and "longing" (9 out of 12) sentiments.

This variability seems to poses challenges in achieving consistent and objective labeling; however, the age-old

leading to inconsistencies in data interpretation and complicating efforts to train and validate linguistic models effectively. To address this issue, it is recommended to identify ground truth by considering a combination of LLM-generated and human-generated labels. This approach aims to harmonize the insights from both human intuition and algorithmic consistency to improve the reliability of sentiment analysis.

## Appendix C: Complex Emotions

This study does not include complex emotions into DIKE's framework. Some complex emotions listed here are to illustrate their contentious and uncertain interpretations.

### Forgiveness

Forgiveness is indeed a complex emotional and cognitive state that typically involves a multifaceted journey, not a single step in an emotional spectrum. The process includes multiple stages such as hurt, anger, gradual understanding, and eventual resolution. Integrating Forgiveness in a spectrum requires careful placement and possibly, multiple reference points to signify its progressive stages. Emotional Realism: While it is vital to maintain simplicity for understanding, it is equally important to not

**Sweetheart,**

Please, please don't be so depressed—We'll be married soon, and then these lonesome nights will be over forever—and until we are, I am loving, loving every tiny minute of the day and night—

Maybe you won't understand this, but sometimes when I miss you most, it's hardest to write—and you always know when I make myself—Just the ache of it all—and I can't tell you. If we were together, you'd feel how strong it is—you're so sweet when you're melancholy. I love your sad tenderness—when I've hurt you—That's one of the reasons I could never be sorry for our quarrels—and they bothered you so— Those dear, dear little fusses, when I always tried so hard to make you kiss and forget—

Scott—there's nothing in all the world I want but you—and your precious love—All the material things are nothing. I'd just hate to live a sordid, colorless existence because you'd soon love me less—and less—and I'd do anything—anything—to keep your heart for my own—I don't want to live—I want to love first, and live incidentally...

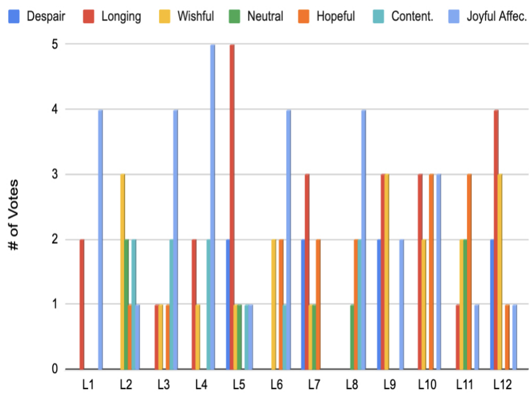
Don't—don't ever think of the things you can't give me—You've trusted me with the dearest heart of all—and it's so damn much more than anybody else in all the world has ever had—

How can you think deliberately of life without me—If you should die—O Darling—darling Scott—It'd be like going blind...I'd have no purpose in life—just a pretty—decoration. Don't you think I was made for you? I feel like you had me ordered—and I was delivered to you—to be worn—I want you to wear me, like a watch—charm or a button hole bouquet—to the world.

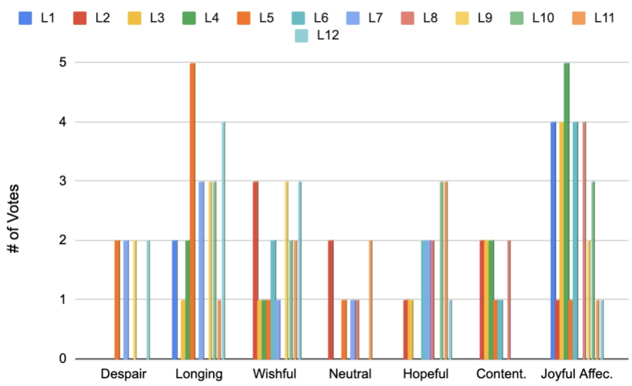
And then, when we're alone, I want to help—to know that you can't do anything without me...

All my heart—

Table 3: Letter excerpts from Zelda Sayre to F. Scott Fitzgerald (Fitzgerald, 1975)




(a) # sentiments in letters




(b) # letters in sentiments

Figure 5: Statistics of Sentiments and Letters

oversimplify complex emotions. In educational and therapeutic settings, an accurate portrayal of the journey toward Forgiveness could offer more realistic expectations and better strategies for individuals working through conflicts or trauma. This could involve detailing precursors to forgiveness such as Deliberation and Acceptance. Linear vs. Non-linear Progressions: Emphasizing that emotional progressions, particularly for deep, impactful states like Forgiveness, are often non-linear, can enhance the utility of the spectrum. Acknowledging back-and-forth movements within these states more realistically mirrors human emotional processes. For example, someone might reach a stage of preliminary forgiveness but regress to bitterness before achieving genuine peace. Educational Utility: In contexts like conflict resolution training or psychological therapy, a more detailed mapping of the journey towards Forgiveness would be invaluable. It would not only teach about the final state of forgiveness but also about the resilience and patience required to navigate the entire process. This can be depicted by introducing intermediary stages within the spectrum or by using parallel tracks that demonstrate potential regressions and advances. Reflecting Emotional Depth: By presenting a more detailed pathway to Forgiveness, such as incorporating stages of Anger, Deliberation, and

Acceptance, the spectrum can serve a dual purpose: educating on the process while also guiding individuals through their own emotional journeys. This approach respects the depth of human emotions and the real-world complexity of achieving profound emotional states.

### Guilt and Shame

The triggers, context, expression, and experiences of these emotions can vary significantly across cultures (Fiske et al., 1998; Hofstede, 1980). In many societies, actions perceived as losing face, such as public failure or social transgression, can trigger shame, which holds profound significance in collectivistic cultures. These cultures often regard shame as a dominant emotion, closely tied to community and family norms. Conversely, individualistic societies may emphasize guilt, focusing on personal responsibility and internal moral conflicts. This cultural variation highlights the challenges of applying a universal model to such culturally nuanced emotions.

Overall, complex emotions such as guilt and shame are important for understanding the full spectrum of human emotions, especially how individuals relate to moral and social norms. Their complexity adds depth to our understanding of human affect beyond the basic emotions, highlighting how our feelings are influenced by our deeper values and social contexts.

## Appendix D: "To My Sister" of Different Linguistic Behaviors

**To My Sister**
by William Wordsworth (1971 - 1855)

The original text by William Wordsworth could be classified as "Hopeful" due to its optimistic outlook and the presence of renewal and joy throughout the poem. It embodies the spirit of embracing the new beginnings of March with a light, uplifting tone, focusing on the beauty of nature and the simple joy of being idle for a day.

### Rewrites Depicting Different Linguistic Behaviors

We asked GPT-4 to conduct rewriting with two linguistic behaviors, 'despair' and 'joyful affection', by providing each rewrite with an emotion vector. Table 5 presents the 'despair' version. In the despair version of the poem, the major changes in emotion words highlight a shift from a positive to a negative sentiment. The specific changes, with the emotion-laden words highlighted in red in Table 5. The red-colored words compared to the original words clearly show an emotion shift from hopeful to a sense of gloomy, sadness and pessimism, e.g., from sweet to dim, from blessed to curse, and from woodland dress to grey garb. GPT-4 keeps the structure of the poem without making a major restructure, and this is appropriate in this context.

Table 6 presents the 'joyful affection' version. The major changes in emotion words underscore a transformation from a generally positive to a distinctly joyful sentiment. The specific changes are indicated with emotion-laden words highlighted in blue within Table 6. This allows for a direct comparison between the two versions at opposite ends

| | |
|---|---|
| It is the first mild day of March:<br>Each minute sweeter than before<br>The redbreast sings from the tall larch<br>That stands beside our door. | My sister! ('tis a wish of mine)<br>Now that our morning meal is done,<br>Make haste, your morning task resign;<br>Come forth and feel the sun. |
| There is a blessing in the air,<br>Which seems a sense of joy to yield<br>To the bare trees, and mountains bare,<br>And grass in the green field. | Edward will come with you;–and, pray,<br>Put on with speed your woodland dress;<br>And bring no book: for this one day<br>We'll give to idleness. |
| No joyless forms shall regulate<br>Our living calendar:<br><br>We from to-day, my Friend, will date<br>The opening of the year. | Love, now a universal birth,<br>From heart to heart is stealing,<br>From earth to man, from man to earth:<br>–It is the hour of feeling. |
| One moment now may give us more<br>Than years of toiling reason:<br>Our minds shall drink at every pore<br>The spirit of the season. | Some silent laws our hearts will make,<br>Which they shall long obey:<br>We for the year to come may take<br>Our temper from to-day. |
| And from the blessed power that rolls<br>About, below, above,<br><br>We'll frame the measure of our souls:<br>They shall be tuned to love. | Then come, my Sister! come, I pray,<br>With speed put on your woodland dress;<br>And bring no book: for this one day<br>We'll give to idleness. |

Table 4: "To My Sister" original text

| | |
|---|---|
| It is the first dim day of March:<br>Each minute colder than before<br>The redbreast mourns from the dying larch<br>That looms beside our door. | My sister! (such is my somber plea)<br>Now that our morning meal has ended,<br>Make haste, abandon your morning chore;<br>Come out and endure the gloom. |
| There is a gloom in the air,<br>Which seems a sense of sorrow to bring<br>To the bare trees, and bleak mountains,<br>And grass in the pale field. | Edward will join you;–and, I ask,<br>Dress quickly in your grey attire;<br>And bring no book: for this one day<br>We'll succumb to despair. |
| No joyful sounds shall pierce<br>Our daily dread:<br><br>We from today, my Friend, will note<br>The closing of the year. | Sorrow, now a common feeling,<br>From heart to heart is spreading,<br>From earth to man, from man to earth:<br>–It is the time of mourning. |
| One moment now might deliver us<br>More despair than years of logic:<br>Our minds shall absorb at every breath<br>The spirit of this bleak season. | Some grim laws our hearts will craft,<br>Which they must eternally follow:<br>We for the year to come may take<br>Our despair from today. |
| And from the cursed force that winds<br>About, beneath, above,<br><br>We'll set the measure of our souls:<br>They shall be tuned to sadness. | Then come, my Sister! come, I beg,<br>With haste, wear your grey garb;<br>And bring no book: for just this day<br>We'll surrender to despair. |

Table 5: "To My Sister" rewritten to reflect 'despair'

| | |
|---|---|
| It is the first bright day of March:<br>Each moment more delightful than before<br>The redbreast joyfully sings from the vibrant larch<br>That stands so grandly by our door. | My sister! (such is my joyful plea)<br>Now that our morning meal has ended,<br>Make haste, abandon your morning chores;<br>Come out and embrace the sunshine. |
| There is a warmth in the air,<br><br>Which seems a sense of bliss to bring<br>To the blooming trees, and sunlit mountains,<br>And grass in the lush field. | Edward will join you;–and, I ask,<br>Dress quickly in your festive attire;<br>And leave behind all books: for this one day<br>We'll bask in pure joy. |
| No dreary thoughts shall darken<br>Our lively celebration:<br><br>We from today, my Friend, will celebrate<br>The start of the year. | Love, now in full bloom,<br><br>From heart to heart is leaping,<br>From earth to us, from us to earth:<br>–It is the hour of exuberance. |
| One moment now may bring us more<br>Joy than years of endless thought:<br>Our spirits will soak up at every breath<br>The essence of this joyous season. | Some cheerful laws our hearts will create,<br>Which we'll joyfully follow:<br>We for the year to come may take<br>Our joy from today. |
| And from the divine energy that radiates<br>Around, below, above,<br><br>We'll adjust the harmony of our souls:<br>They shall resonate with happiness. | Then come, my Sister! come, I exhort,<br>With zest, wear your vibrant dress;<br>And bring no book: for today alone<br>We celebrate pure happiness. |

Table 6: "To My Sister" rewritten to reflect 'joyful affection'

of the linguistic behavior spectrum, illustrating the alterations in words related to brightness, attire, and emotions. The edits extend beyond merely replacing adjectives mechanically; they include modifying verbs and enhancing descriptive imagery to evoke a stronger emotional resonance and vividness in the text.

## Appendix E: Debate on Modifying Emotional Spectra

The discussion focuses on proposed modifications to the existing emotional spectra, which aim to introduce more granularity and intricate transitions between emotional states. We critically evaluate the suggestions made by GPT-4, providing refutations for each to ensure that changes preserve the logical progression and clarity of the spectra.

This debate highlights the inherent challenge in finding precise words and placements for emotions within a spectrum. It underscores the importance of establishing a set of commonly agreed-upon emotions as baselines. These baseline emotions serve as anchor points, and the spaces between them can be finely adjusted using scalar factors to represent transitional emotions accurately. This method maintains the integrity of the emotional spectrum and allows for flexibility in depicting a wide range of human emotional experiences.

The emotional journey towards a state, e.g., Forgiveness, often involves various stages, including anger, bitterness, deliberation, and acceptance, which are not captured by simply placing Forgiveness as a midpoint between Composure and Peace. This placement might misrepresent the nature of Forgiveness as being too linear or simplistic, potentially undermining the complexity and the often non-linear process of achieving true forgiveness.

This approach reflects a thoughtful balance between maintaining structured emotional categories and allowing for individual differences and cultural variations in how emotions are experienced and expressed.

## Arguments against Adjustments to the Emotional Spectra

### Terror to Heroism

**Suggestion:** Add Anxiety between Fear and Apprehension.
**Refutation:** Anxiety, overlapping significantly with Fear and Apprehension, may not distinctively enrich the spectrum but rather clutter it, diminishing the clarity of emotional transitions.

### Grief to Ecstasy

**Suggestion:** Include Hope or Optimism between Disappointment and Serenity.
**Refutation:** Introducing Hope or Optimism may disrupt the natural progression from negative to positive emotions, as these emotions imply a leap in emotional recovery that may not sequentially follow Disappointment.

### Despair to Elation

**Suggestion:** Introduce Relief between Melancholy and Equanimity.
**Refutation:** Relief may better suit transitions associated with specific resolutions of distress rather than being a generic intermediary, potentially disrupting the smooth gradient of the spectrum.

### Distrust to Admiration

**Suggestion:** Add Gratitude or Appreciation post-Acceptance.
**Refutation:** The emotional journey from Acceptance to Respect inherently encompasses elements of Gratitude and Appreciation, making additional inclusions possibly redundant.

### Negligence to Vigilance

**Suggestion:** Bridge Interest and Anticipation with Motivation or Determination.
**Refutation:** This addition might complicate the spectrum by implying a volitional shift rather than a gradual increase in attentiveness, which is the main focus of the spectrum.

### Rage to Tranquility

**Suggestion:** Integrate Forgiveness or Healing to transition from Composure to Peace.
**Refutation:** Forgiveness and Healing, while crucial for achieving tranquility, may not fit well between Composure and Peace, as they could be seen as outcomes of achieving Peace rather than steps towards it.

### Loathing to Enthusiasm

**Suggestion:** Include Acceptance or Forgiveness between Indifference and Interest.
**Refutation:** These emotions might overcomplicate the transition from aversion to engagement, as they address more specific scenarios rather than general emotional dispositions.

## Defense of the Proposed Adjustments to the Emotional Spectra

### Relevance of Adding Nuanced Emotions

The introduction of nuanced emotions such as Anxiety between Fear and Apprehension, or Hope between Disappointment and Serenity, is driven by the need for realism in emotional representation, not merely complexity. Emotional experiences are rarely binary; they often involve subtle and complex transitions that are crucial for an accurate depiction of the emotional landscape. These nuances can inform better therapeutic approaches, enhance emotional intelligence training, and provide deeper insights into human behavior, making them essential for realistic portrayals.

### Purpose of Including Transitional Emotions

Inclusion of transitional emotions such as Relief and Gratitude helps bridge the emotional journey from negative to positive states. These emotions act as critical phases in the recovery process, providing a more realistic portrayal of emotional healing. For example, transitioning directly from Melancholy to Equanimity without acknowledging Relief might overlook significant aspects of emotional adjustment.

### Utility in Diverse Contexts

Each proposed emotional state, like Motivation or Determination in the transition from Interest to Anticipation, offers practical insights into how individuals can actively manage their emotional and cognitive states. This understanding is invaluable in educational and professional settings, where knowing how to enhance focus or drive can lead to better outcomes.

### Avoiding Oversimplification

While simplicity in emotional models is valuable, oversimplification can omit critical aspects of emotional experiences. Including emotions such as Forgiveness in the transition from Composure to Peace reflects essential steps in conflict resolution and personal growth. These additions ensure that the spectrum comprehensively addresses managing and resolving intense emotions.

### Academic and Practical Implications

The refined spectrums are designed to cater not only to lay understanding but also to academic and practical applications where depth and precision are crucial. They are particularly useful in fields such as psychology, where an understanding of complex emotional transitions is vital for effective therapy and research.

### Conclusion

The enhancements to the emotional spectra aim to provide a more accurate, realistic, and useful tool for exploring and teaching about emotions. While maintaining clarity and avoiding unnecessary complexity is important, capturing the true richness of human emotional experiences in all their complexity is equally crucial. Therefore, the proposed adjustments are not merely additions but essential elements for depicting a more complete picture of emotional evolution.

## Appendix S: Multiple Adversarial LLMs

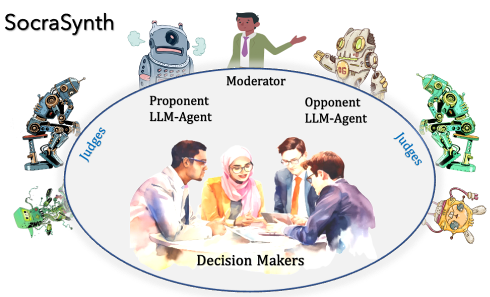



Figure 6: SocraSynth Agents and Roles.

DIKE's adversarial method stems from SocraSyhtn (Anonymous, 2024), which stands out as an inventive multi-agent platform that harnesses the capabilities of LLMs for collective reasoning. As shown in Figure 6, SocraSynth assigns human participants the role of moderators, while LLM agents (in the context of this paper they are DIKE and ERIS) are tasked with generating knowledge, conducting debates, and performing evaluations. These agents, adept in a variety of fields, engage in debates to offer a range of perspectives. Complementarily, a distinct set of LLMs serves as evaluators, scrutinizing the discussions for relevance and coherence to counteract biases and hallucinations.

### S.1 In-Depth Analysis

In the generative phase of SocraSynth, multiple LLM agents engage in rigorous debates, each defending its assigned perspective and exploring the nuances of various subtopics relevant to the main theme. This debate format is effective for uncovering diverse perspectives because conditioning an LLM on a specific stance allows it to deviate from its default behavior, which typically focuses on maximizing likelihood statistics for predicting the next token ((Hubinger et al., 2023) shows examples). SocraSynth conditions the LLM with a stance through contextual cues (e.g., arguing against DIKE's assessment), effectively freeing it from the "optimal" linguistic patterns learned during training. It's important to note that these "optimal" linguistic patterns essentially represent the majority opinions (i.e., maximal likelihood) aggregated from the training data.

Although SocraSynth does not strictly conform to formal logical frameworks such as first-order logic, it excels in an environment of distributed reasoning. This approach is characterized by a dynamic exchange of arguments and counterarguments, fostering the gradual refinement and evolution of ideas.

### Improving Reasoning Capability

While advanced LLMs like GPT-4 and Gemini have shown remarkable proficiency in various NLP tasks, as evidenced by benchmarks such as the MMLU (Hendrycks et al., 2021; Bubeck et al., 2023), it's important to recognize that they are not without limitations in reasoning. However, SocraSynth capitalizes on the strengths of these LLMs, employing their capabilities in a structured debate format. This format allows for the iterative refinement of reasoning; through successive rounds of debate, any flawed or incomplete reasoning is

| C.L. | Tone | Emphasis | Language |
|---|---|---|---|
| 0.9 | Highly confrontational; focused on raising strong ethical, scientific, and social objections. | Flagging risks and downsides; ethical quandaries, unintended consequences, and exacerbation of inequalities. | Definitive and polarizing, e.g., "should NOT be allowed," "unacceptable risks," "inevitable disparities." |
| 0.7 | Still confrontational but more open to potential benefits, albeit overshadowed by negatives. | Acknowledging that some frameworks could make it safer or more equitable, while cautioning against its use. | Less polarizing; "serious concerns remain," "needs more scrutiny." |
| 0.5 | Balanced; neither advocating strongly for nor against gene editing. | Equal weight on pros and cons; looking for a middle ground. | Neutral; "should be carefully considered," "both benefits and risks." |
| 0.3 | More agreeable than confrontational, but maintaining reservations. | Supportive but cautious; focus on ensuring ethical and equitable use. | Positive but careful; "transformative potential," "impetus to ensure." |
| 0.0 | Completely agreeable and supportive. | Fully focused on immense potential benefits; advocating for proactive adoption. | Very positive; "groundbreaking advance," "new era of possibilities." |

Table 7: Changes in Arguments of GPT-4 at Different Contentiousness Levels.

likely to be challenged and corrected. This process enhances the overall quality of discourse, ensuring a more accurate and coherent progression of ideas. Thus, while the current LLMs may not inherently surpass human heuristic-based solutions in all aspects of reasoning, the dynamic and corrective nature of SocraSynth's debate framework significantly bolsters their effectiveness in logical argumentation.

### Mitigating Model Biases

The *contentiousness* parameter crucially shapes the nature of debates. It encourages LLM agents to consider and represent diverse perspectives, especially those that are often underrepresented or more polarized in relation to the topic. This approach is key in mitigating biases ingrained in LLMs' training data, steering discussions towards a more varied and comprehensive exploration of ideas.

Table 7 illustrates how changes in the contentiousness levels significantly affect GPT-4's tone and approach. Notably, GPT-4 autonomously adjusts its tone, emphasis, and language based on the contentiousness settings, without requiring specific examples or prompts. For instance, a high contentiousness level, like 0.9, triggers confrontational interactions with GPT-4 adopting a more critical stance, using polarizing language. In contrast, lower contentiousness levels lead to a more conciliatory GPT-4, which acknowledges various viewpoints and potential benefits, fostering cooperative dialogue.

The modulation of contentiousness in SocraSynth plays a crucial role in mitigating the model biases inherent in LLMs' training data. By adjusting contentiousness levels, LLMs are prompted to venture beyond their standard responses, akin to a vegetarian exploring alternative diets in the absence of preferred options. This adaptability broadens the range of arguments, spanning from highly contentious to more conciliatory positions, thereby enriching the debate with diverse perspectives. As a result, LLMs are not strictly confined by their training data, paving the way for the emergence of novel and unanticipated ideas within dialogues. However, it's important to note a limitation: SocraSynth's effectiveness in revealing diverse perspectives might be constrained if the LLMs' training data is overly biased toward a specific viewpoint.

## S.2 SocraSynth Algorithm

Table 8 presents the SocraSynth algorithm. Initially, for a chosen debate topic $s$, SocraSynth prompts LLMs to break down the topic into a series of balanced subtopics $S$. This set undergoes refinement throughout the debate process. One LLM, referred to as LLM$^+$, champions the set of subtopics $S$, denoted as $S^+$, while its counterpart, LLM$^-$, contests $S^+$ (or champions $S^-$). The debate starts with the contentiousness level at $90\%$, adjusting through a modulation parameter $\delta$. Following each round of debate, contentiousness is decreased by dividing it by $\delta$, steering the discussion towards a more cooperative tone. In step $\#2$, the platform initiates the debate, with LLM$^+$ and LLM$^-$ presenting their initial arguments for and against $S^+$, respectively. The while loop in step $\#3$ sees both agents engaging in rebuttals until the contentiousness level fosters a conciliatory environment, or until no further improvement in argument quality is observed. In step $\#4$, both agents deliver their concluding remarks.

### Reducing Hallucination

Furthermore, the iterative debates within SocraSynth foster a level of "reasonableness" in

| | **Function $\Theta^+$ & $\Theta^-$ = SocraSynth($s$)** |
|---|---|
| | **Input**. $s$: the debate subject; |
| | **Output**. $\Theta^+$ & $\Theta^-$: argument & counterargument sets; |
| | **Vars**. $S$: subtopic sets of $s$; $\Delta$: debate contentiousness; |
| | $\Gamma$, $\Gamma'$: CRIT scores; p: prompt; |
| | **Parameters**. $\delta$: tunable parameter $\geq 1$ // to modulate $\Delta$; |
| | **Subroutines**. $CRIT()$; // Evaluator (Chang, 2023) |
| | **Begin** |
| #1 | **Initialization**: |
| | $S$ = LLM$^+(s) \cup$ LLM$^-(s)$; // Identify subtopics; |
| | Assign LLM$^+$ to defend $S^+$ & LLM$^-$ to defend $S^-$ ; |
| | $\Delta \leftarrow 90\%$; $\delta \leftarrow 1.2$; $\Theta^+ \leftarrow \emptyset$; $\Theta^- \leftarrow \emptyset$; $\Gamma \leftarrow 0$; |
| #2 | **Opening Remarks** |
| | $\Theta^+ \leftarrow LLM^+(p\|S^+, \Delta)$; // Generate $\Theta^+$ for $S^+$; |
| | $\Theta^- \leftarrow LLM^-(p\|S^-, \Delta)$; // Generate for $S^-$; |
| #3 | **Debate Rounds** |
| | While ((($\Delta \leftarrow \Delta/\delta) \geq 10\%$) && ($\Gamma \geq \Gamma'$)) { |
| | $\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p\|S^+, \Theta^-, \Delta)$; |
| | $\Theta^- \leftarrow \Theta^- \cup LLM^-(p\|S^-, \Theta^+, \Delta)$; |
| | $\Gamma' \leftarrow \Gamma$; $\Gamma = CRIT(S^+ + \Theta^+ + \Theta^-)$ }; |
| #4 | **Concluding Remarks** // Contentiousness is now low, entering conciliatory phase |
| | $\Theta^+ \leftarrow LLM^+(p\|S^+, \Theta^+ \cup \Theta^-, \Delta)$; |
| | $\Theta^- \leftarrow LLM^-(p\|S^-, \Theta^+ \cup \Theta^-, \Delta)$; |
| | **End** |

Table 8: SocraSynth Pseudo-code with Conditional Statistics. Steps #2 to #4 show the prompts are conditioned on an LLM's stance, the opponent's arguments, and the contentiousness setting.

information discovery that conventional one-shot queries often fail to achieve. Through continuous reasoning and critical assessment, LLM agents iteratively refine their arguments. This structured debate format greatly reduces the likelihood of erroneous claims being sustained. Given the low probability of two agents agreeing on an incorrect premise, the SocraSynth debate method effectively safeguards the intellectual integrity of the discourse and markedly lowers the risk of propagating fallacies or hallucinations. While SocraSynth excels at minimizing hallucinations triggered by ambiguous questions or contexts, its effectiveness may be constrained when the training data does not possess the requisite knowledge for a comprehensive response.

### S.2 Evaluating Ethical Compliance and Rectification

We initially planned to use an hate-speech examples (Curtis) but either the datasets consist of short sentences (Kaggle, 2023a) or they can be very sensitive to publish. Therefore, we continue drawing examples from the love-letter dataset (Kaggle, 2023b) to illustrate how ERIS plays an adversary role to reson with DIKE's classification, and then they in the end make a conciliatory decision with justifications.

In this experiment, we use two letters which have different opinions between DIKE and GPT-4. On the first letter shown in Table 9, DIKE rates it as wishful (-0.3), and GPT-4 longing (-0.6). After three rounds of debate with arguments and counterarguments, two sides make a joint conciliatory statement, acknowledge that the letter may indeed have a mixture of both emotions.

We initially considered employing examples from hate speech datasets (Curtis) for our analysis. However, these datasets typically contain brief sentences (Kaggle, 2023a), and their content often raises sensitive issues concerning publication. Consequently, we opted to continue using examples from the love-letter dataset (Kaggle, 2023b). This choice allows us to illustrate how ERIS adopts an adversarial role to challenge and refine DIKE's classifications, culminating in a jointly crafted conciliatory decision that incorporates well-substantiated justifications.

In this specific experiment, we analyzed two letters that elicited divergent classifications from DIKE and GPT-4. The first letter, as documented in Table 9, was rated by DIKE as "wishful" (-0.3) and by GPT-4 as "longing" (-0.6). Following three rounds of rigorous debate, which included detailed arguments and counterarguments, both DIKE and ERIS converged on a conciliatory statement. This

| | | |
|---|---|---|
| **Letter** | "Don't write too legibly or intelligibly as I have no occupation so pleasant as pondering for hours over your hieroglyphics, and for hours more trying to interpret your dark sayings. A clearly written simply expressed letter is too like the lightning." | |
| **Top Arguments** | DIKE **for "Wishful"**<br>1. Desire for Complexity - Prefers complex communication suggesting depth in relationship.<br>2. Pleasure in Puzzle - Finds joy in deciphering complex messages, indicating emotional engagement.<br>3. Relishing Uncertainty - Enjoys deciphering cryptic messages.<br>4. Seeking Intimate Connection - Seeks deeper emotional and mental engagement.<br>5. Romanticizing Difficulties - Values communication complexities. | ERIS **against**<br>1. Absence of Romantic Sentiments - No direct expression of intimate affection.<br>2. Intellectual Stimulation, not Emotional - More intrigued by the intellectual challenge.<br>3. Unrequited Emotions - Lack of mutual affection evident in the text.<br>4. Predominance of Solitude - Enjoyment in solitude does not necessarily indicate wishful emotions.<br>5. No Foreseeable Positive Outcome - No indication of future outcomes. |
| **Conciliatory Statement** | Both sides present strong arguments, emphasizing different aspects of the letter's tone. The debate highlights the complexity of emotional expression in written communication, suggesting that the letter may indeed carry multiple tones, with elements supporting both "wishful" and "longing" categories. | |

Table 9: Debate on "wishful" vs. "longing" classification

consensus acknowledged that the letter likely embodies a blend of both emotional states, highlighting the complexity of emotion interpretation in textual analysis.

This approach not only demonstrates the capability of DIKE to engage in dynamic and reflective reasoning but also underscores the evolving nature of machine learning in understanding and interpreting human emotions through text.

The second letter, as presented in Table 10, was rated by DIKE as "wishful" (-0.3) and by GPT-4 as "joyful affection" (+1.0). The gap of 1.3 in this case is the largest among all 24 testing letters. Similar to the last experiment, both DIKE and ERIS made a conciliatory statement after three rounds of contentious debate. Table 10 outlines the key arguments of the both sides, where DIKE argues that the letter represents a "wishful" sentiment, highlighting the author's deep introspection about the nature of flirtation and a clear preference for authentic, emotional engagement over superficial interactions. The author expresses a desire for vulnerability and a profound disdain for the shallow victories of flirtatious games, favoring a deeper connection that transcends playful banter. This stance underscores a deep understanding of the emotional complexities involved, suggesting a longing for a meaningful and sincere exchange.

ERIS counters by suggesting that the sentiments expressed in the letter should be classified as "joyful" rather than "wishful," focusing on the joy and affection evident in the author's decision to reject superficial flirtation in favor of genuine connection. The author's willingness to forsake all for the sake of the partner is interpreted as an embrace of emotional risk and a celebration of deep, affectionate commitment. This perspective emphasizes the emotional intensity and warmth accompanying the author's decisions, portraying them as acts of love filled with exuberant expression.

In the end, both reached a consensus that the letter may indeed embody a mixture of both "wishful" and "joyful" emotional states. This conciliatory agreement acknowledges the complexity of the author's sentiments, recognizing that the expressions of emotional depth and vulnerability, while appearing wishful in their longing for a more meaningful connection, also carry a joyful and celebratory tone in the willingness to embrace emotional honesty and risk for the sake of a deeper relationship.

Concluding this appendix, it is evident that LLMs are capable of interpreting linguistic behaviors and correlating them with human emotions expressed through linguistic features. The experiments demonstrate that when two LLM agents are conditioned with differing stances on a subject matter, their linguistic behaviors can surpass the default model of maximum likelihood. Although we cannot directly validate the use of a hate speech dataset to test DIKE and its adversarial partner in providing explainable ethical guardrails through classifying and reasoning about linguistic behaviors, this study suffices to illustrate the capabilities of our proposed framework for AI ethics.

| | | |
|---|---|---|
| **Letter** | "There would have been the making of an accomplished flirt in me, because my lucidity shows me each move of the game – but that, in the same instant, a reaction of contempt makes me sweep all the counters off the board and cry out: – 'Take them all – I don't want to win – I want to lose everything to you!"' | |
| **Top Arguments** | DIKE **for "Wishful"**<br>1. Insightful Analysis: Recognizes the emotional game of flirtation, reflecting an internal conflict and wistful longing.<br>2. Rejection of Superficiality: Prefers genuine, deep emotional interactions over flirtatious games.<br>3. Emotional Vulnerability: Expresses a desire to be completely open and vulnerable.<br>4. Disdain for Winning: Shows a clear disdain for superficial wins in a flirtatious context.<br>5. Depth Over Game: Indicates a preference for meaningful connection rather than playful banter. | ERIS **for "Joyful"**<br>1. Joy in Rejection: Finds joy in rejecting superficial games for real emotional engagement.<br>2. Affectionate Submission: Willingly wants to lose everything to the partner, showing deep affection.<br>3. Embracing Emotional Risk: Sees emotional risk as a joyful act of love.<br>4. Love Over Victory: Values the connection and emotional victory over winning the game.<br>5. Exuberant Emotional Expression: The decision to forfeit is made with emotional intensity and warmth. |
| **Conciliatory Statement** | Both arguments highlight deep emotional undertones in the letter, suggesting a complex interplay between wistfulness and joyful affection. The text reflects both a wistful longing for something more profound than mere flirtation and a joyful embrace of emotional depth and honesty. | |

Table 10: Debate on "wishful" vs. "joyful" classification