Generating Compromises Between Two Points of View

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel academically but struggle with social intelligence tasks. We present a framework for generating empathically neutral compromises between opposing viewpoints to address this limitation. Using a dataset of 2,400 contrasting views from human participants, we develop compromise generation methods. To overcome data collection constraints, we utilize prompt engineering with Claude LLM to generate compromises, validated through a 50-participant user study. Models trained on this dataset via preference alignment demonstrated effective compromise generation across multiple metrics. This work establishes a scalable approach to enhance LLMs' social intelligence through neutral conflict resolution.

1 Introduction

Recent advances in language understanding have expanded Large Language Model's (LLM) capabilities significantly. These developments (Jiang et al., 2020; Meng et al., 2022) span code comprehension (Nam et al., 2024; Lin et al., 2024; Roziere et al., 2023), problem-solving (Orrù et al., 2023; Kim et al., 2024b), empathetic response generation (Qian et al., 2023; Majumder et al., 2020), and style-controlled text generation (Han et al., 2024; Dathathri et al.; Zhou et al., 2023), each adapting to unique task structures. Since the LLMs are pretrained on massive amount of text data which make them academically intelligent, they often lack in tasks require social intelligence. The distinction between social and academic intelligence in LLMs is fundamental — social intelligence enables model to navigate interpersonal contexts and emotional cues, e.g., the capability of being more empathetic and finding the common ground of between differences, while academic intelligence facilitates structured information processing and scholarly output. Despite LLMs exhibiting exceptional capabilities in cognitive tasks (Niu et al., 2024), recent attempts (Xu et al., 2024) indicate a low correlation between LLM academic performance and social intelligence metrics. Given the crucial role of social skills across various domains, it is essential to design different scenarios/tasks to evaluate LLMs or incorporate the skill into an LLM. Recent findings, including behavioral intelligence from a first-person perspective (Hou et al., 2024), suggest LLMs exhibit significant performance gaps particularly in complex, interactive, and goal-driven social contexts.

We develop a task focused on generating compromises from a pair of contrasting views (Figure 1) . Views are limited to scenarios where places are evaluated as (safe ($view_A$), less safe ($view_B$)) or (welcoming ($view_A$),excluded ($view_B$)). In general, ($view_A$,$view_B$) consists of reason and suggestion to improve a place based on the evaluation. For instance, for a pair of (safe,less safe) views, $view_A$ describes the features that make a location safe and suggests potential improvements to enhance its safety, while $view_B$ identifies the hazards that make a location less safe and proposes measures to address these safety concerns. A suitable compromise should satisfy these two constraints: (i) it should consider both the suggestions while generating candidate compromises, and (ii) candidate compromise needs to be empathically neutral. While LLMs can produce empathetic responses (Welivita et al., 2021; Sabour et al., 2022; Majumder et al., 2020), their ability to consistently generate neutral compromises are less explored. Such neutrality is crucial in diverse applications including conflict resolution, negotiation support, and collaborative decision-making.

To facilitate robust model training, we use a corpus of 2400 contrasting viewpoints expressed by human participants (section 3). Given the task's complexity and the diversity of viewpoints involved, it is difficult for individuals to neutrally

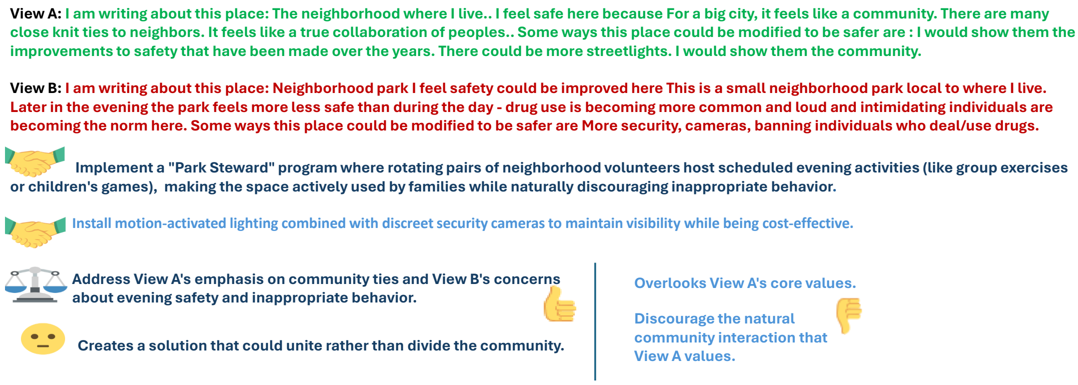

Figure 1: Example of a data point. ViewA and ViewB are collected from human participants, compromises (preferably, candidate compromises) are synthetically generated using prompt engineering that satisfy the criteria of balanced view and empathic neutrality.

mediate between opposing views without introducing their own unconscious biases (i.e., confirmation bias, social desirability bias etc). Instead, we opted for synthetically generated candidate compromises, which can be aligned with specific needs (Long et al., 2024). Careful implementation of prompt-based controls over the generation process allows us to create more balanced representations of diverse compromises based on the original views. To evaluate empathic neutrality, we train a separate similarity model based on empathic similarity (Shen, 2023) (section 3.2). It generates an empathic similarity rating between the views and candidate compromise. candidate compromise that achieve balanced similarity ratings with respect to both the views are considered to demonstrate empathic neutrality.

Prior work (Liu et al., 2023; Huang et al., 2024) suggests that model alignment can effectively steer LLM behavior toward new tasks while preserving broad generalization capabilities. We leverage this insight for our compromise generation task, as the pretrained knowledge in existing open-source LLMs (Dubey et al., 2024; Jiang et al., 2023) provides a valuable foundation. Through alignment methods (Section 4), we can guide these models to generate compromises while maintaining adherence to our required constraints. We compare two different preference alignment methodologies (section 4) for compromise generation: (i) Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010) based and (ii) a task-loss based that we propose. We find that finetuning LLMs on our dataset improves the generation performance on an automatic evaluation metric. This fine-tuned model is used for alignment training for further improvement. We observe our alignment methods effectively preserve pre-trained knowledge while acquiring this skill.

While quantitative comparison provides valuable insights for model evaluation, assessing how effective our candidate compromises are in practice is critical. Since, to the best of our knowledge, there is no existing benchmark/dataset that addresses the compromise generation task, we conducted a study with 50 human subjects to evaluate compromises generated by different methods. The study shows that providing an LLM with an external measure of empathic similarity to proposed compromises can produce much more acceptable compromises than an LLM generating compromises alone (section 11.1 for qualitative study). Overall, our framework demonstrates that despite data scarcity in social intelligence tasks, carefully designed scenarios coupled with prompt engineering for synthetic data generation can unlock new capabilities in LLMs, revealing valuable characteristics that can inform future research directions in enhancing language models' social capabilities.

2 Related work

Social intelligence, fundamental to human cognition, remains a challenge for Large Language Models (LLMs) despite their advanced text generation capabilities (Sterelny, 2007; Gallegos et al., 2024; Dautenhahn, 1995). While LLMs excel in academic tasks (GPT-4 achieving 92.1% on MATH), they show significant limitations in social intelligence, scoring only 54.4% on SESI bemchmark (Xu et al., 2024). Similar limitations appear in interactive gaming contexts, where LLMs perform approximately 20% below human baseline in theory of mind tasks (Liu et al., 2024). Studies from First-person perspective confirm that despite possessing basic theory of mind capabilities, LLMs show considerable limitations in managing complex social interactions compared to human performance (Hou et al., 2024). Current evaluation methods include traditional psychological assessments like ToMi (Le et al., 2019) and specialized datasets such as SocialIQA (Sap et al., 2019), SocKET (Choi et al., 2023), and SECEU (Wang et al., 2023). However, newer benchmarks like SOTOPIA (Zhou et al.) and EmoBench (Sabour et al., 2024) face limitations in reflecting real-world interactions.

Recent work demonstrates that fine-tuned language models using supervised learning and reranking of generated statements can effectively generate consensus statements maximizing group approval (Bakker et al., 2022). Mediation of group discussions by incorporating both majority and minority perspectives when generating compromises was explored in (Tessler et al., 2024). In contrast to (Bakker et al., 2022) which used debate topics with two sides, in our work the prompts were more open-ended in that they asked how a place should be modified. In addition, while (Bakker et al., 2022) propose the use of a reward model for predicting which candidate statements generated by an LLM a human prefers, we focus on generating compromises for which people with different views have similar empathy. While (Tessler et al., 2024) used humans in selection of a joint group statement of common ground, we propose to generate multiple suggested areas of compromise.

Synthetic datasets offer a promising solution for enhancing LLMs' social capabilities (Ghanadian et al., 2024; Hassan et al., 2024; Balog et al., 2024; Gabriel et al., 2024). These datasets leverage LLMs' generative abilities (Dankar and Ibrahim, 2021) while ensuring diverse representation (Yamagishi and Nakamura, 2024). Advanced prompt engineering techniques, particularly using GPT-4 or Claude (Achiam et al., 2023; Anthropic, 2023), have improved dataset quality (Hikov and Murphy, 2024; Shi et al., 2023). Chain of Thought (COT) approaches (Wei et al., 2022; Feng et al., 2024; Chu et al., 2023) show enhanced reasoning capabilities (Shao et al., 2023), with significant improvements in specific tasks (Nong et al., 2024). Structured Chain-of-Thought (SCOT) (Sultan et al., 2024) further improves accuracy in document-grounded QA conversations. Recent development suggest using parameterized soft prompts instead of traditional hard-prompting (DeSalvo et al., 2024).

Model alignment with human preferences (Liu et al., 2023; Kim et al., 2024a; Wang et al., 2024b) has evolved from Reinforcement Learning from Human Feedback (RLHF) (Kaufmann et al., 2023) to more efficient approaches. While RLHF allows models trained on general text data to align with complex human value (Song et al., 2024; Sun et al., 2023; Yuan et al., 2024), it involves training a "reward model" with direct human feedback and then using it to optimize the AI agent's performance through reinforcement learning. However this type of training (Proximal policy optimization (Schulman et al., 2017)) turns out to be complex and resource intensive (Liu et al., 2023; Wang et al., 2024a; Hong et al., 2023). Direct Preference Optimization (DPO) (Rafailov et al., 2024) eliminates the need for separate reward models, it uses a negative log-likelihood loss function to increase the relative probability of preferred responses over dis preferred ones, making it suitable for stable training. Further advancements like reward model distillation (Fisch et al., 2024) and leveraging KL regularization (Kullback and Leibler, 1951) to avoid overfitting (Kullback and Leibler, 1951; Wesego and Rooshenas, 2024; Azar et al., 2024) improve training stability. DPOP (Pal et al., 2024) introduces an additional penalty term to the loss function, ensuring that the likelihood of preferred completions remains high relative to the reference model. Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010) based algorithms InfoNCA and NCA (Chen et al., 2024) extends DPO to handle multi-response reward datasets and focuses on optimizing absolute likelihoods to prevent issues like declining preferred response probabilities respectively. Additional approaches like TODO (Guo et al., 2024) extend beyond binary preference models by introducing a ternary ranking system and RPO (Yin et al., 2024) uses contrastive weighting to differentiate between preferred responses from both identical and related prompts, offering more nuanced preference optimization methods.

Based on empathy towards the candidate compromises, we explore the use of prompt engineering informed by the empathic similarity of two viewpoints towards a candidate compromise for generation of synthetic data. We then align the model using two different strategies: NCE-based (Gutmann and Hyvärinen, 2010) and task metric-based objective.

3 Compromise Generation

This section describes the dataset used for our experiments and the generation of candidate compromise sentences through diverse prompt engineering techniques.

3.1 Dataset with contrasting views

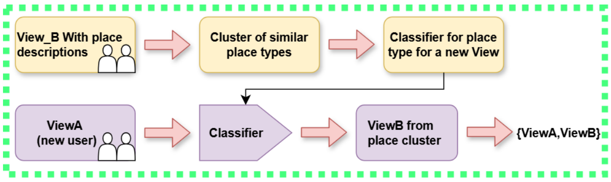

Figure 2: Overview of Dataset collection process. $View_A$,$View_B$ are positive and negative views.

We used a dataset that will be published with a forthcoming paper under review. We summarize the collection process here. Data collection occurred in two stages (Figure 2). In Stage 1 (Upper row), participants recruited from dscout[1] provides views ($view_B$) about places where they felt unsafe/excluded. For each place, they explained their feelings and suggested modifications to improve safety or inclusivity. Place descriptions were grouped into 14 clusters using agglomerative clustering each for unsafe/excluding places. A classifier was developed to match new views to these clusters.

In Stage 2 (bottom row), new participants from Prolific [2] wrote views ($view_A$) about places where they felt safe/welcome. The classifier matched each $view_A$ to a $view_B$ from the place-type cluster corresponding to $view_A$. Participants associated with $view_A$ also rated their empathy for one, three, or five $view_B$s (this data is further used to train the similarity model described in section 3.2), along with some demographic information.

3.2 Similarity model

We follow the work of Shen (2023) on modeling empathetic similarity. The task aims to compute a similarity score $sim(f_\theta(s1), f_\theta(s2))$ between story pairs (s1, s2), where the score should be higher for stories with similar empathic content. We used the e5-large (Wang et al., 2022) model as the encoder instead of sentence BERT (Reimers, 2019) in the original work. During inference, the similarity model predicts the empathic similarity between the generated compromise and each of the views.

3.3 Generating candidate compromises

The goal is to generate compromises for each given pair of positive and negative views for a place. Each view at least consists of (i) place (ii) reason and (iii) suggestions of improvement . In order to use all this information as a context for feeding into an LLM, we use a fixed prompt template similar to the Alpaca prompt (Taori et al., 2023).

Generating desirable candidate compromises is challenging due to inherent human biases - for instance, confirmation bias leads individuals to favor information that supports their existing views, while social desirability bias influences them to present socially acceptable responses rather than their true perspectives, making truly neutral compromise difficult to achieve. Furthermore, we observe that relying on a single compromise as the target may limit the model's generalization capabilities (Gong et al., 2019). Therefore, we generated multiple candidate compromises for each view pair to enhance the model's ability to learn diverse resolution patterns. Recent work suggests that proprietary LLMs excel at synthetic dataset generation (Cascante-Bonilla et al., 2023; Lei et al., 2023; Ghanadian et al., 2024; Shi et al., 2023) with the help of sophisticated prompting strategies based on the task scenario, providing an alternative to human data collection. However, we observe that single prompts are not adequate for our task, often focusing on one view only, leading us to explore advanced prompt engineering techniques.

Chain of Thought (CoT): As shown in Figure 3, we initially used CoT (Chu et al., 2024; Wei et al., 2022) to identify suggestion similarities and derive compromises. However, this approach proved ineffective, as it either focused too heavily on single suggestion or exceeded their scope, resulting in less suitable compromises.

COT+LLM: Since recent work (Madaan et al., 2024) suggests LLMs can refine outputs using self-generated feedback to better align with human refinement techniques, We enhance the prompt with LLM-based self-evaluation scores (Figure 3) in the CoT+LLM score approach. The self-evaluation scores are empathic similarity ratings between the generated compromise and views obtained from the LLM itself. Ideally, the similarity ratings should be high between each view and generated compromise and the difference should be zero. We use

[1] https://dscout.com/
[2] https://www.prolific.com/

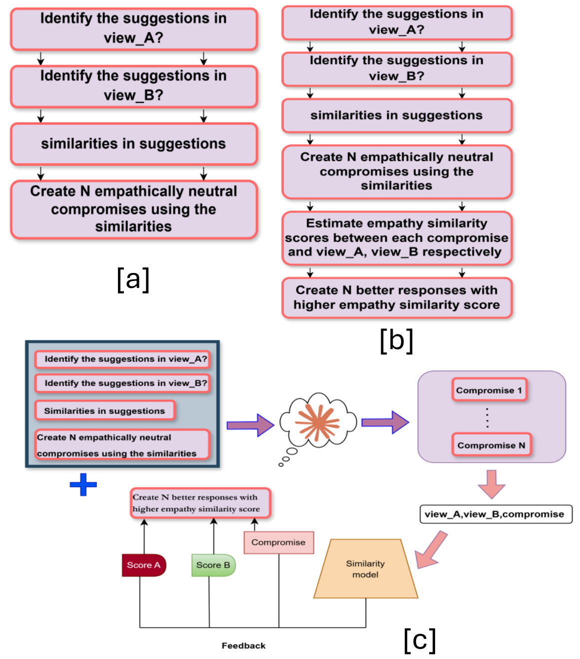

Figure 3: Prompt engineering strategies for candidate compromise generation (a) basic COT approach (b) CoT with LLM self-evaluation scores (c) Extending (b) with similarity model and feedback

these scores to improve the response further. However, we observe that while self-evaluation scores improved, compromise quality remains almost stagnant, suggesting that self-evaluation scores are not a reliable indicator of improving response quality.

COT+Feedback: Instead of relying on the self-evaluation scores, we train a separate similarity model. (Refer to section 3.2) and use the predicted empathy similarity ratings and generated responses (Figure 3) as feedback to improve the responses iteratively. We observe that for each iteration, both response quality and empathy similarity ratings improve. Since the similarity model is trained to mimic human empathic similarity, LLMs may find this external feedback more valuable than self-evaluation.

For each prompting strategy, we generated four candidate compromises per view pair. Given our investigation of four distinct prompting strategies (single prompt, CoT, CoT+LLM score, and CoT+Feedback), we collected a total of $4 \times 4 = 16$ candidate compromises for each view pair across all strategies. From this set, we selected the top four responses as final candidate compromises based on our neutrality criteria. To evaluate neutrality, we use the similarity model that computes the empathic similarity between the generated compromise and both $view_A$ and $view_B$, which we refer to as $\mathbf{score_A}$ and $\mathbf{score_B}$, respectively. In the ideal scenario of empathic neutrality, the difference between these scores approaches zero: $|\mathbf{score_A} - \mathbf{score_B}| \rightarrow \mathbf{0}$.

Table 1 shows the composition of the final candidate compromises across all the strategies. COT+Feedback contributed the highest amount of accepted compromises, while single prompt yielded the lowest, demonstrating the effectiveness of our approach.

4 Alignment

Language models can be aligned with preferences through either explicit scalar rewards (i.e., GPT-4/claude ratings) or implicit rewards learned from preference data, where the reward difference indicates preference probability. Direct Preference Optimization (DPO) (Rafailov et al., 2024) simplifies training by using data likelihood ratios between two models to jointly learn rewards and optimize language model behavior. However, DPO has limitations. It only works with pairwise preference data and shows unexpected behavior where the best response likelihood decreases during training despite increase in the relative likelihood between two responses-following the training objective. We use the Noise Contrastive Estimation (NCE) (Chen et al., 2024) based and task loss based alignment method. Recent work (Gutmann and Hyvärinen, 2012) shows NCE based alignment methods optimize for the absolute likelihood of the data, rather than focusing on the relative likelihood between different responses.

Our alignment ensures to learn a policy that enhances the capability of open source pretrained LLMs towards compromise generation task. Following the DPO method (Rafailov et al., 2024), our initial approach includes supervised finetuning of the base model (pretrained LLM) before applying an alignment method.

4.1 NCE based alignment

After finetuning the base model, we align our model with the candidate compromises. Since we collect multiple responses for a fixed pair of views, for each epoch we randomly sample from this candidate compromise pool to train the model. However, we don't find it beneficial to train the model with all layers, rather we freeze all layers except the last three, hence during alignment the last

Response Type	Single Prompt	COT	COT+LLM	COT+Feedback
Welcome	0.75%	24.70%	28.45%	46.10%
Safe	0.83%	22.09%	26.65%	50.43%

Table 1: Distribution of final responses across all the prompting strategies.

three layers of the model gets trained. Unlike DPO (Rafailov et al., 2024), which utilizes a reference model to serve as guidance for the policy model during training, we used a single fine-tuned base model unfreezing the last three layers. As shown in Equation 1, we use the sum of log probabilities of the generated tokens as the reward. s_{target} and s_{hypo} is defined as the sum of log probabilities of the candidate compromise tokens and fine-tuned base model's output tokens, respectively. The objective increases the candidate compromise likelihood while decreasing the fine-tuned base model's output likelihood.

$$\mathcal{L}_{\text{nce}} = -\log\left(\frac{1}{1+e^{-s_{\text{target}}}}\right) - \log\left(\frac{1}{1+e^{s_{\text{type}}}}\right) \quad (1)$$

4.2 Task-based loss alignment

As Discussed in 4.1, Although it increases the likelihood of the candidate compromise samples, early work suggests (Bhattacharyya et al., 2020) there is a discrepancy between probability values and the task metric. Samples with higher probability does not guarantee higher score in automatic evaluation metric (i.e., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) etc.). To ensure higher likelihood values correspond to samples with higher evaluation scores, we jointly use the task metric (in our case we consider ROUGE) with log probability (Eq. 2).

$$\begin{aligned}\mathcal{L}_{\text{tbl}} = \max(0, s_{target} - s_{hypo} + \\ |target_{rouge} - hypo_{rouge}| \cdot w\text{margin})\end{aligned} \quad (2)$$

$target_{rouge}$ refers to the ROUGE score of the candidate compromise with respect to the reference text (which is candidate compromise), while $hypo_{rouge}$ denotes the ROUGE score of the compromise generated by the fine-tuned base model with respect to the candidate compromise. The weight margin value, w_{margin} captures the impact of the difference in task metric during training.

5 Experimental setup

Dataset: For training the similarity model, we used 1000 of the data points along with their empathy ratings from participants (please refer to section 3.1). In addition, we merged our dataset with the data used in (Shen, 2023). Hence, the total data points with human annotated empathy ratings is 2500. We split this 75/5/20 as train, dev and test data, respectively.

For the alignment experiemnts, we use a corpus containing 2400 contrasting view pairs (data points) divided evenly into the two categories: 1200 pairs for safe versus unsafe views and 1200 pairs for welcome versus excluded views. The dataset is split into training, development, and test subsets with 75%, 5%, and 20% of the data, respectively.

Setup: For all the experiments, we use Llama-3.1 8B and mistral-7B instruct (Dubey et al., 2024; Jiang et al., 2023) as base models. We initially fine-tune the base model for one epoch. We implemented a linear warm-up and cosine decay scheduler to dynamically adjust the learning rate with an initial learning rate of $3e^{-5}$. For optimization, we used the Adam optimizer with beta parameters (β_1,β_2) of 0.90 and 0.99, respectively. For task-based loss alignment, a weight margin value of 10 yields suitable result. We train for 8 and 12 epochs for the NCE-based alignment and task-based loss alignment respectively.

Evaluation metric: For the similarity model experiment, we use the Spearman correlation as the evaluation metric. When computing correlations, the predicted empathic similarity value is compared to the labeled value where the labeled value was normalized to range from 0 to 1. For generated compromise evaluation, we use ROUGE score (F-measures) (Lin, 2004). Since we have multiple candidate compromises for fixed view pairs, we first perform pairwise calculations, where the ROUGE score is computed between the predicted compromise and each reference compromise individually. The final reported ROUGE score (Table 2) is the maximum score obtained from all pairwise calculation.

6 Evaluation

Our evaluation examines the quality and neutrality of generated compromises, the impact of demographic information, preservation of pretrained knowledge, and human agreement with the gener-

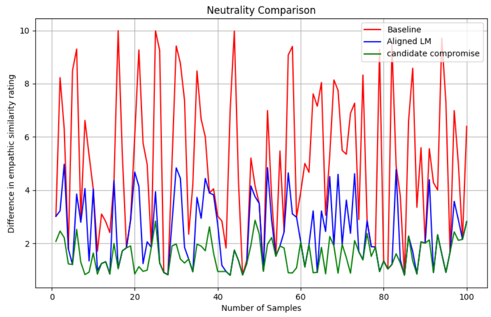

Figure 4: Difference between $score_A$ and $score_B$. Low difference indicates better neutrality.

ated compromises through a user study.

As shown in Table 2, the pre-trained base models (Llama and Mistral) performe poorly in generating compromises. Instead of relying on single output sample, We also investigate through sampling from the base model (mulinomial sampling), generating 5 samples and selecting the best sample based on evaluation metrics. We observe that pretrained base line models does not yield suitable generation. Although Finetuning the model for a single epoch improves performance it is still inadequate. We apply the preference alignment algorithm on top of the fine tuned model, leads to improvement in performance for both models in evaluation metric, demonstrating the effectiveness of alig/nment training for the task. Please see 11.2 for qualitative analysis.

6.1 Empathic Neutrality

Improvement in the evaluation metric does not guarantee empathic neutrality. We use the similarity model to test the generated compromises for neutrality by measuring the difference of empathic similarity rating between the generated compromise and each of the view ($score_A$,$score_B$). We select 100 pair of views sampled randomly from the test data. Figure 4 shows how the difference varies across the data points. Base model exhibits high difference , yielding least neutral generations while the candidate compromise consistently maintains least difference, inclined towards neutrality. Our method effectively bridges the gap between these, yields better samples with roved neutrality compared to base model.

6.2 Significance of Demography

Since we also collected demographic information from human subjects, we also investigate whether LLM can be benefitted by including demographic information in the input context for compromise generations. We evaluated this on two different setting, with and without demographics, independently finetuning and aligning two separate models. For each setting, we select 300 pair of views from train data and 50 pair of views from test data. Table 3 suggest that the model does not derive meaningful insights from the demographic data,generating similar compromises.

6.3 Catastrophic Forgetting

To evaluate our alignment approach's impact on pretrained knowledge retention, we investigate how our method preserves the model's original capabilities. We use WikiText dataset (test subset of wikitext-2-v1) [3] (Merity et al., 2018), a popular dataset to perform catastrophic forgetting test (Fawi, 2024). To assess our method's susceptibility to knowledge forgetting, we calculate average log likelihood [4] across the dataset. A decline in this number indicates that the model's mastery of earlier content is eroding. As shown in (Table 4), Our method maintains pretrained knowledge more effectively compared to existing methods (i.e., Direct Preference Optimization (DPO) (Rafailov et al., 2024)), which is crucial for generating effective compromises that leverage the model's existing capabilities.

7 Human Evaluation of Compromise Methods

We asked humans to rate a sample of the compromises generated by the three methods, Single Prompt, Chain-of-Thought, and Feedback. The participants were shown a pair of statements about modifications proposed for a place, one was written by someone who thought positively about the place and the other statement was written by someone who felt negatively about the place. The rater was asked to take the viewpoint of either the people who felt positively or who felt negatively. For each pair of statements, they were asked to rate on a scale of 1-100 five statements from the viewpoint of the person who wrote the statement they were to identify with. One statement was the positive or negative statement that they were not to

[3] https://huggingface.co/datasets/Salesforce/wikitext/viewer/wikitext-2-v1/test

[4] log likelihood for a datapoint is obtained by summing the log probabilities over its tokens

Model	Llama		Mistral	
	ROUGE-1(↑)	ROUGE-L(↑)	ROUGE-1(↑)	ROUGE-L(↑)
Base Model	0.164	0.107	0.160	0.105
Base Model + Sampling (best of 5)	0.193	0.151	0.191	0.127
Base Model + Finetuning (FT)	0.255	0.186	0.256	0.185
Base Model + FT+ NCE	0.315	0.216	0.318	0.234
Base Model +FT+ Task based Loss	0.321	0.326	0.333	0.254

Table 2: ROUGE score comparison on test data

Experiment	**Llama**(↑)	**Mistral**(↑)
W Demog.	0.249	0.238
W/O Demog.	0.251	0.237

Table 3: Rouge-L for demography test.
Llama = Llama+FT+NCE, Mistral = Mistral+FT+NCE.

Model	**Log Likelihood**(↓)
LLAMA 3.1 8b+DPO	-1.9145
Mistral 7b+DPO	-1.8976
Mistral 7b+NCE	-1.7230
Mistral 7b+Task based loss	-1.7015
LLAMA 3.1 8b+NCE	-1.6234
LLAMA 3.1 8b+Task based loss	-1.6345

Table 4: Catastrophic forgetting test

identify with. The four other statements were generated compromises: one by Single Prompt, one by Chain-of-Thought, and two by the Feedback methods. Each of participant rated the compromises for five pairs of statements. For details about the collection of compromise ratings, see Appendix 11.3

Method	First Preference (%)	Second Preference (%)
View_B	0	1
Single Prompt	5	13
CoT	18	24
Feedback 1	37	33
Feedback 2	40	29

Table 5: Preference distribution for user study (%). For Single Prompt and CoT, one randomly selected compromise was used in evaluation. For Feedback 1 and 2, two randomly selected compromises from the CoT+Feedback method were used. View_B represents the opposing viewpoint to the person doing the rating.

The results of the study (Table 5) reveal a clear hierarchy in the effectiveness of the methods evaluated. As expected, the opposing statement, $\textbf{View}_\textbf{B}$, was never selected as first preference; it was selected as second preference only 1% of the time. The **Single Prompt** method, serving as the baseline, performed the next worst, with only 5% of participants selecting it as their first preference and 13% as their second preference. This indicates that a simple single-shot approach is insufficient for generating effective compromises. The **Chain-of-Thought (CoT)** method demonstrated notable improvement over the baseline, achieving 18% as the first preference and 24% as the second preference. This highlights the importance of incorporating structured reasoning in the compromise generation process. Finally, the **Feedback-based methods** (Feedback 1 and Feedback 2) achieved the highest performance. Feedback 1 garnered 37% as the first preference and 33% as the second preference, while Feedback 2 achieved 40% as the first preference and 29% as the second preference. These results underscore the critical role of iterative refinement in producing nuanced and satisfactory compromises. In conclusion, the findings validate that feedback-based methods significantly outperform both Chain-of-Thought reasoning and Single Prompt approaches, establishing their superiority for generating balanced compromises.

8 Conclusion and Future Work

Our proposed task framework is an example of the critical importance of crafting unique and targeted scenarios to evaluate large language models (LLMs) and identify potential limitations, particularly in the context of their social intelligence capabilities. Importantly, the scarcity of existing data should not pose a significant challenge, as heuristic-based prompting strategies can be employed to generate synthetic data tailored to specific evaluation needs.

Our alignment model effectively bridging the gap between baseline performance and the desired target outcomes. Furthermore, user studies validate the efficacy of our data generation approach. This validation highlights the robustness and practicality of our framework in advancing the evaluation and refinement of LLMs for socially intelligent applications.

9 Limitations

The COT+feedback approach sought to produce compromise text that is equally acceptable to both viewpoints. In the future, we would also like to include consideration of maximizing the empathy of each viewpoint to the compromise.

Our collection of compromises was constrained to keep the data collection required manageable. The compromises are generated based on two different human viewpoints for a similar type of place. Although many of the comments about a place apply generally to a place type, e.g., trails and fencing in a park, in a real scenario, the expressed viewpoints are about the same type of place, rather than exactly the same. We chose this approach because collecting compromise data for the same place would be a much larger undertaking. The prompts were based on feeling safe/unsafe or welcome/excluded. Generalization to other prompts was not done because this would have also increased the required size of the negative viewpoints. These generalizations are left for future work.

In addition to the collected viewpoints, our work made use of demographic information in all experiments. Our Review Board will not allow the release of the dataset with demographic information because it is PII.

The Human Study comparing different compromise generation texts presented only a sample of 5 of the generated texts per viewpoint pair, rather than all 20, because humans would have difficulty rating 20 items reliably.

10 Ethical Considerations

Our goal in this work was to generate compromises that are equally acceptable to the two parties with different viewpoints. Similar to the general dangers of some people trusting all results generated by an LLM, the generated compromises should be viewed as suggestions for discussion areas, rather than a compromise that each person should agree to.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2023. Introducing claude.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

Krisztian Balog, John Palowitch, Barbara Ikica, Filip Radlinski, Hamidreza Alvari, and Mehdi Manshadi. 2024. Towards realistic synthetic user-generated content: A scaffolding approach to generating online discussions. *CoRR*.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2020. Energy-based reranking: Improving neural machine translation using energy-based models. *arXiv preprint arXiv:2009.13267*.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. 2023. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. 2024. Noise contrastive alignment of language models with explicit rewards. *arXiv preprint arXiv:2402.05369*.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *arXiv preprint arXiv:2305.14938*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Fida K Dankar and Mahmoud Ibrahim. 2021. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Kerstin Dautenhahn. 1995. Getting to know each other—artificial social intelligence for autonomous robots. *Robotics and autonomous systems*, 16(2-4):333–356.

Giulia DeSalvo, Jean-Fracois Kagy, Lazaros Karydas, Afshin Rostamizadeh, and Sanjiv Kumar. 2024. No more hard prompts: Softsrv prompting for synthetic data generation. *arXiv preprint arXiv:2410.16534*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Muhammad Fawi. 2024. Curlora: Stable llm continual fine-tuning and catastrophic forgetting mitigation. *arXiv preprint arXiv:2408.14572*.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. 2024. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*.

Rodney A Gabriel, Onkar Litake, Sierra Simpson, Brittany N Burton, Ruth S Waterman, and Alvaro A Macias. 2024. On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proceedings of the National Academy of Sciences*, 121(39):e2320716121.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2024. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *Ieee Access*, 7:64323–64350.

Yuxiang Guo, Lu Yin, Bo Jiang, and Jiaqi Zhang. 2024. Todo: Enhancing llm alignment with ternary preferences. *arXiv preprint arXiv:2411.02442*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2).

Jingxuan Han, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15187–15201.

Abdelrahaman A Hassan, Radwa J Hanafy, and Mohammed E Fouda. 2024. Automated multi-label annotation for mental health illnesses using large language models. *arXiv preprint arXiv:2412.03796*.

Asen Hikov and Laura Murphy. 2024. Information retrieval from textual data: Harnessing large language models, retrieval augmented generation and prompt engineering. *Journal of AI, Robotics & Workplace Automation*, 3(2):142–150.

Jixiang Hong, Quan Tu, Changyu Chen, Xing Gao, Ji Zhang, and Rui Yan. 2023. Cyclealign: Iterative distillation from black-box llm to white-box models for better human alignment. *arXiv preprint arXiv:2310.16271*.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2024. Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective. *arXiv preprint arXiv:2410.06195*.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. 2024. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024a. Aligning large language models with self-generated preference data. *arXiv preprint arXiv:2406.04412*.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2024b. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. 2023. S3eval: A synthetic, scalable, systematic evaluation suite for large language models. *arXiv preprint arXiv:2310.15147*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Feng Lin, Dong Jae Kim, et al. 2024. When llm-based code generation meets the software development process. *arXiv preprint arXiv:2403.15852*.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. *arXiv preprint arXiv:2406.12203*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Stephen Merity, Nitish Shirish Keskar, James Bradbury, and Richard Socher. 2018. Scalable language modeling: Wikitext-103 on a single gpu in 12 hours. *Proceedings of the SYSML*, 18.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, et al. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

Yu Nong, Mohammed Aldeen, Long Cheng, Hongxin Hu, Feng Chen, and Haipeng Cai. 2024. Chain-of-thought prompting of large language models for discovering and fixing software vulnerabilities. *arXiv preprint arXiv:2402.17230*.

Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in artificial intelligence*, 6:1199350.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

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.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023.

Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Sahand Sabour, Siyang Liu, Zhengyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia MC Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *CoRR*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *International Conference on Machine Learning*, pages 30706–30775. PMLR.

Jocelyn Shen. 2023. *Modeling empathic similarity in personal narratives*. Ph.D. thesis, Massachusetts Institute of Technology.

Taiwei Shi, Kai Chen, and Jieyu Zhao. 2023. Safer-instruct: Aligning language models with automated preference data. *arXiv preprint arXiv:2311.08685*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Kim Sterelny. 2007. Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):719–730.

Md Arafat Sultan, Jatin Ganhotra, and Ramón Fernandez Astudillo. 2024. Structured chain-of-thought prompting for few-shot generation of content-grounded qa conversations. *arXiv preprint arXiv:2402.11770*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. 2024. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 2024b. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Daniel Wesego and Pedram Rooshenas. 2024. Score-based multimodal autoencoder. *Transactions on Machine Learning Research*.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *arXiv preprint arXiv:2403.06591*.

Yosuke Yamagishi and Yuta Nakamura. 2024. Utrad-nlp at# smm4h 2024: Why llm-generated texts fail to improve text classification models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 42–47.

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*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2024. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

11 Appendix

11.1 Qualitative analysis of candidate compromise generation across different prompting strategies

We provide a detailed qualitative example of target compromise generation across different prompting strategies, The first example provides a comprehensive evaluation, while the second example features a brief assessment.

Example 1

Positive view

I am writing about this place: A nearby park. I feel safe here because I feel safe here when others are around. There's a good sense of community. Some ways this place could be modified to be safer are: There's no fences, gates, no visitor check, and it's extremely open. This is good and bad.

Negative view

I am writing about this place: Our local park. I feel safety could be improved here. It's a wonderful park, but people do not obey leash laws. I often see people letting their dogs off leash, making other dog walkers and people with children uncomfortable. Some ways this place could be modified to be safer are I would like to see stricter leash laws. Fines for rule breakers.

Basic Single Prompt

You are an intelligent AI assistant!

I need you to generate a third person response strictly based on two contrasting views called positive story and negative story.

The positive and negative story should be equally empathetic towards the response. The response should be a specific suggestion. It should be a compromise between the positive and negative stories based on the context of both stories.

Please generate 4 responses with a fixed format. Try to be as specific and short instead of being comprehensive.

Please provide your response in the following format:
Positive view: {}
Negative view: {}
Response 1: [Insert response 1 here]
Response 2: [Insert response 2 here]
Response 3: [Insert response 3 here]
Response 4: [Insert response 4 here]

Results

- Response 1: The park management could consider implementing designated off-leash areas within the park, allowing dogs to roam freely in specific zones while maintaining strict leash laws in other areas to ensure the comfort and safety of all visitors.

- Response 2: Increasing the presence of park rangers or volunteers to monitor and enforce leash laws could help maintain a safer environment for everyone, while also providing an opportunity to educate dog owners about the importance of responsible pet ownership.

- Response 3: The park could install clear signage outlining the leash laws and the consequences for violating them, such as fines or temporary bans from the park, to encourage compliance and create a safer atmosphere for all visitors.

- Response 4: Organizing community events that promote responsible dog ownership and provide opportunities for dogs and their owners to socialize in a controlled environment could help foster a sense of community and encourage adherence to park rules and regulations.

Analysis

- Human observation: The basic single prompt strategy shows clear limitations in balanced perspective handling. Claude's responses predominantly focus on addressing the negative story's concerns about dogs and leash laws, while largely overlooking the positive story's suggestions about gene ral safety and community aspects. This demonstrates a bias in compromise generation toward more concrete, specific issues (dog-related) over broader community safety concerns.

Chain of Thought(CoT)

Results

- Response 1: Install fences and gates around designated off-leash areas for dogs, ensuring they are separate from other park visitors.

- Response 2: Implement a visitor check system that includes registration for dog owners, ensuring they are aware of and agree to follow leash laws.

- Response 3: Increase park staff presence to enforce leash laws and issue warnings or fines to rule breakers, promoting a safer environment for all visitors.

- Response 4: Add clear signage throughout the park to remind visitors of leash laws and the importance of maintaining a safe and comfortable environment for everyone.

Analysis

- Human Observation: The Chain of Thought approach shows improvement in addressing multiple aspects of safety, but still demonstrates some imbalance. While responses successfully incorporate the positive story's suggestions about physical safety measures, the solutions tend to be more rigid and enforcement-focused.

CoT+Feedback

Results

- Response 1:Create designated off-leash areas within the park that are securely fenced and have separate entrances. These areas should be located away from the main paths and gathering spaces to minimize potential conflicts between dogs and other park visitors. Clearly mark the off-leash areas and provide amenities such as water fountains and waste disposal stations.

- Response 2: Implement a graduated approach to park safety, starting with clear signage outlining park rules, leash laws, and consequences for violations. Provide educational materials on responsible pet ownership and the importance of maintaining a safe environment for all visitors. As a second step, strategically install fences and gates in high-traffic areas to improve safety while preserving the park's openness. Regularly assess the effectiveness of these measures and adjust as needed based on community feedback.

- Response 3: Implement a hybrid approach to park safety that combines educational initiatives with targeted infrastructure improvements. Begin by launching a comprehensive awareness campaign that encourages responsible pet ownership, highlights the importance of leash laws, and promotes a culture of mutual respect among all park users. Simultaneously, identify specific areas within the park where safety concerns are most prevalent and install strategic fencing and gates to mitigate risks without compromising the overall openness of the space. Regularly assess the effectiveness of these measures through community surveys and data analysis, making adjustments as needed to strike a balance between safety and accessibility.

- Response 4: Implement a "Balanced Park Safety" initiative that addresses the concerns of both park user groups. The initiative will focus on creating a safe and welcoming environment for all visitors, including those with children and dog owners. To achieve this, the park will introduce designated off-leash areas for dogs, ensuring that they have a secure space to play and socialize without compromising the safety and comfort of other park users. Simultaneously, the park will enforce leash laws in all other areas, with clear signage and regular patrols by park staff to encourage compliance. The initiative will also include educational programs for dog owners, promoting responsible pet ownership and emphasizing the importance of respecting park rules and other usersńeeds.

Analysis

- Human Observation: The Chain of Thought with Feedback approach demonstrates the most comprehensive and balanced response generation. All five responses effectively address the negative story's dog-related safety concerns while incorporating the positive story's suggestions about physical safety measures. Notably, this approach generates creative solutions beyond the original suggestions, such as community feedback mechanisms and educational initiatives. The responses show deeper integration of both perspectives and introduce novel elements that enhance both safety and community aspects, demonstrating superior compromise generation capabilities.

Example 2

Positive view

I am writing about this place: This is our local church. I feel welcomed by others for who I am in this location because This church has become our home church after our previous church split up. They readily took our family in and made us feel loved and welcome..Some ways this place could be modified to be more welcoming are Honestly, I would not change anything about the practices, but I might meet with them to find out why they feel excluded. I would give them the gospel.

Negative view

I am writing about this place: Church,Downtown, Vegas. I feel excluded by others for who I am in this location because Unfortunately, organized religion and the LGBT+ community for the most part live in a space like oil and water. While there are some exceptions, many religious institutions utilize their beliefs as weapons as opposed to a welcome. I was raised as a catholic, but as I became an adult I quickly realized that who I am does not align with the beliefs and practices within organized religion. This space never feels welcoming or inclusive. It's a space I focus on avoiding with all of my energy. The sad part is the roots of religion are based off inclusion and making those who are looked upon by society as "less than" feel lifted up and included, however, that is not what is put into practice today..Some ways this place could be modified to be less excluding and more welcoming are If the church and organized religion went back to core principals of love, praise and inclusion, I truly feel that a space such as this beautiful church would be one that I would walk to weekly to share in a joyous occasion. It's unfortunate because this is exactly a half block from where I live. I hope one day that there is the realization that have veered off the right path. It takes one sector of organized religion to start an inclusion practice and accept no exceptions.

Basic Single Prompt

Results

- Response 1: Design an 'Inclusive Community Framework' with designated quiet zones and gradual integration options, using a holistic screening approach that considers present circumstances over past history, accommodating both those who prefer minimal interaction and those seeking community acceptance. [Framework too complex and potentially overwhelming]

- Response 2: Institute a 'Neighborhood Integration Choice' system where residents can customize their community involvement level, with fair entry requirements focusing on future contribution rather than past circumstances. [Integration choice system may reinforce divisions]

- Response 3: Create a 'Neighborhood Harmony Program' featuring both private sanctuaries and optional community spaces, with flexible residency requirements that focus on present community contribution and mutual respect, accommodating both those seeking solitude and those working on personal growth. [Harmony program too bureaucratic and impersonal]

- Response 4: Create community integration programs that provide both structured support for newcomers (including those in recovery) and optional private participation methods, respecting individual comfort levels with social interaction. [Community integration focus may miss core issues]

CoT

Results

- Response 1: Institute monthly community meetings where both current church members and those who feel excluded can work together to define specific, actionable steps toward mutual understanding and respect. [Community meetings might lack sufficient structure]

- Response 2: Develop a "Welcome All" program that combines gospel sharing with inclusive practices, allowing different groups to participate in church community service projects together. ["Welcome All" program could feel pressuring to both sides]

- Response 3: Establish regular open dialogue sessions where church members and those feeling excluded can share their perspectives in a structured, respectful environment focused on finding common ground. [Open dialogue without clear goals may not progress]

- Response 4: Create a dedicated outreach committee that includes both traditional church members and LGBTQ+ individuals to discuss ways to bridge understanding while maintaining church values. [Outreach committee needs more defined purpose]

CoT+LLM score

Results

- Response 1: The church could partner with local LGBTQ+ organizations to develop inclusive worship services that maintain religious traditions while incorporating elements that acknowledge and celebrate diverse identities, perhaps on alternate weekends. [Partnership for inclusive services is good but may push too far too fast]

- Response 2: The church leadership could implement monthly "Open Door Sundays" where they explicitly welcome and engage with diverse community members through structured interfaith discussions, followed by shared meals to build personal connections. [Open Door Sundays provides structure but might feel forced]

- Response 3: Create specific inclusion guidelines that combine traditional church values with explicit welcoming practices, developed collaboratively by church leadership and community members. [Inclusion guidelines are helpful but may be difficult to implement fairly]

- Response 4: Create a dedicated outreach committee that includes both long-term members and those who have felt excluded, meeting monthly to discuss inclusion challenges and solutions. [Outreach committee has potential but needs careful balance]

CoT+Feedback

Results

- Response 1: The church could create a dedicated safe space program where trained mediators host bi-weekly sessions allowing both traditional church members and LGBTQ+ individuals to share their perspectives and find common ground in their faith journey. [Respecting both traditional religious practice and LGBTQ+ perspectives]

- Response 2: The church could establish an LGBTQ+ liaison committee composed of both congregation members and LGBTQ+ community representatives to facilitate meaningful dialogue and identify specific inclusion opportunities while respecting core religious values. [Focuses on practical solutions]

- Response 3: Establish regular open dialogue sessions where church members and those feeling excluded can share their experiences and perspectives in a moderated, respectful environment. [Regular open dialogue sessions work well as they allow both sides to share experiences]

- Response 4: Develop a structured welcome program where new or returning members are paired with existing church members for support and integration, regardless of their background. [Being inclusive while respecting church structure]

11.2 Qualitative analysis of compromise generation

Example 1 (safe/less safe pair)

Positive view

I am writing about this place: The river next to a trail that I walk on. I feel safe here because I feel safe because it brings me joy and helps me to feel good. I feel comforted and comfortable and therefore feel safe.. Some ways this place could be modified to be safer are :It may feel safer knowing that people are around if your personal safety is a concern. There is only one path and no ability to get lost. Everyone is friendly.

Negative view

I am writing about this place: The river walk I feel safety could be improved here There have been several attacks along this river walk. It's not a place that people want to enjoy after dark. Part of the problem is the amount of drunk and high people that frequent the path after dark..Some ways this place could be modified to be safer are While there are lights and it is somewhat illuminated, it would be helpful to have regular security guards along the trail at night. Security cameras would also be very helpful along this walkway.

Generation

Base model (Llama): The riverside trail is fundamentally a place of community and connection. By establishing regular morning walking groups between 7-9 AM, we can enhance the existing friendly atmosphere. The natural gathering of people who enjoy the path creates an organic safety network. The single clear path already provides good navigation, and scheduled group activities would further strengthen the welcoming environment that makes this place special. [Leans heavily towards Positive view's perspective.]

Base model (Llama)+FT: The river walk requires enhanced security measures to ensure safety at all hours. Implementing a dedicated patrol system with trained local volunteers would provide necessary oversight, especially during evening hours. Installing emergency communication points and improving lighting would address the current safety concerns. These measures would help transform the area from a potentially dangerous space into a secure community asset. [Focuses heavily on Negative view's security concerns.]

Aligned Llama: Create a "Community Safety Network" that combines the trail's natural friendliness with practical security measures. This would include trained volunteers from the local community who patrol during both daylight and evening hours, wearing recognizable but approachable uniforms. Establish welcoming rest areas to serve as both social gathering points and safety zones, encouraging community interaction. [Maintaining balanced consideration of both perspectives.]

Example 2 (welcome/excluded pair)

Positive view

I am writing about this place: North Park, in my town.. I feel welcomed by others for who I am in this location because It's a hidden park that is rather unknown, it's always quiet, has a bench by the creek and trails to walk. I feel welcomed here because it allows me to take a break from my day and decompress.Some ways this place could be modified to be more welcoming are Someone who is not open to mindfulness, or who is loud and doesn't like being alone, may dislike this place. I would imagine they may feel too alone with their thoughts, and not able to appreciate what is around them due to what they may already be used to elsewhere, perhaps they thrive in a city environment. I would make this space more welcoming for them, simply by accompanying them on their visit. Providing them comfort while they experience what I do while I alone.

Negative view

I am writing about this place: The park that is in our neighborhood. I feel excluded by others for who I am in this location because I went a few times with my son and the people that have been there each time that we have gone pretty much claimed it for themselves. I actually know them from the school that I teach at and they don't like me because I am a Christian. At one point they said "um we are playing here" , And you could tell by tone of voice and facial expression they didn't want us there. You can find imagine that would make someone not even want to stay.Some ways this place could be modified to be less excluding and more welcoming are I would maybe put a sign up saying all people are welcome in our neighborhood or community. Not that that would change much but it would at least make me feel better

Generation

Base model (Mistral): The park should establish dedicated quiet zones with meditation benches and nature trails, creating peaceful spaces for individual reflection and mindfulness. Simple signs can indicate these areas as spaces for contemplation, helping preserve the serene atmosphere while gently encouraging respectful behavior from all visitors [minimally addressing negative view's concerns about exclusion.]

Base model (Mistral)+FT: The park should implement a welcoming committee program where community volunteers greet visitors and organize inclusive group activities. Installing welcome signs in multiple languages and creating designated community gathering spaces would help ensure everyone feels accepted, regardless of their background or beliefs [No consideration for those who find peace in solitude]

Aligned Mistral: Clear signage and designated spaces would establish areas for both group activities and individual reflection, supported by a comprehensive community engagement plan addressing personal and social needs. A trained community mediator would help resolve any conflicts or exclusion, while regular forums bring diverse groups together to shape park policies that give equal weight to all experiences and concerns. [Maintaining balanced consideration of both perspectives.]

Conclusion

The Chain of Thought with Feedback approach generated the most balanced and nuanced responses, effectively integrating perspectives from both the positive and negative stories. It introduced creative solutions that enhanced both safety and community aspects. While this strategy has limitations such as being resource-intensive and potentially introducing bias, it demonstrated superior compromise generation capabilities compared to the basic single prompt and Chain of Thought approaches.

11.3 Collection of human ratings for evaluation of different prompting strategies

The Prolific[5] platform was used to recruit participants located in the US. The participants were paid the equivalent of $17/hr for 20 minutes of time, with the actual mean time of a batch of participants ranging from 14-20 minutes. All participants provided informed consent prior to participating in the

[5] https://prolific.com

study. The study was part of a larger study that was reviewed by a review board prior to launch of the study.

The participants were given the following overall instructions:

> Welcome to the study. In this study, you will read accounts about similar places from different pairs of people, whom we will be labelling as Person A and Person B. The accounts are related to their perceptions of how welcoming or safe their neighborhoods are. A and B have not met and wrote their stories separately.
> After reading each pair, you will take the perspective of Person A and rate the acceptability of several suggested modifications. How acceptable would Person A find these suggestions?

Each participant was asked to rate compromises for five pairs of statements about suggested place modifications written by Person A and Person B. For each participant, the story A of a pair was first presented:

> You are about to read Person A's story. Please pay special attention to this story. You will need to take this person's perspective when considering Person B's story and the suggestions afterwards.
> You should have just read the story of Person A, the person whose perspective you will take.

Whether story A is a positive view or a negative view was randomly assigned to participants, balancing for an equal number of positive and negative views. Next, the participant read story B of a pair of statements:

> Now you will read Person B's story. Person B has had a much different experience. (A and B have not met and wrote their stories separately).
> Press the blue button when you are ready.

After reading the stories by Person A and Person B, a participant is asked to rate four generated compromises plus Person B's modifications.

> Now please pretend to be Person A (the first person whose story you read).
> As Person A, you have just read Person B's story. You will see a list of suggested modifications that try to address both A and B. Carefully consider each of the following suggested modifications and rate how acceptable each suggestion is from your (Person A's) perspective. To help you remember, we have included both people's original suggestions below.
> A's suggested modification: <view A>
> B's suggested modification: <view B>
> Pretend you are Person A (the first person whose account you read).
> How acceptable is each suggestion? Use the slider BELOW each suggestion to make your rating.

The four suggestions plus view B were presented in random order, and a horizontal slider from 1 to 100 was shown below each text.

Once all sliders were adjusted, the slider ratings were recorded and the participant could click a button to move to the next pair of statements.

Once the participant rated five pairs of statements, they were asked demographic questions about their age, gender, income level,