Generating Compromises Between Two Points of View

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models often falter when asked to reconcile opposing views, revealing biases toward asymmetric compromises. We study whether they can instead produce compromises that respects both narratives without favoring either side. Given paired accounts about the same place type with contrasting perspectives, we generate multiple compromise candidates using prompt strategies guided by an empathy-oriented similarity model and select those with balanced similarity to both inputs. Evaluation by 50 participants on a subset of a corpus of 2,400 paired views showed that single prompt compromises tended to be biased towards one view while use of external neutrality evaluation during feedback-guided prompting yields more neutral compromises over single-prompt and chain-of-thought (CoT) baselines. We also examine whether alignment enables models to acquire neutrality heuristics despite initial bias. This task offers a compact probe of how models integrate conflicting cues and exhibit neutrality in social reasoning.

1 Introduction

The ability to compromise is essential in navigating situations that involve divergent viewpoints. Beyond its social utility, compromise generation provides a way to study how models balance conflicting signals and whether they approximate human like heuristics of fairness and neutrality. Studying whether LLMs exhibit these behaviors provides insight into their capacity for socially intelligent reasoning [22, 8, 7]. Prior studies have suggested that while LLMs can generate consensus statements from diverse preferences [2] or facilitate deliberation by reducing polarization [25], their outputs often lack consistent neutrality, failing to equitably represent opposing sides without implicit biases or heuristics that favor one narrative.

Despite their strong performance on academic benchmarks, LLMs show persistent weaknesses in tasks requiring social intelligence [15], such as empathy and impartial mediation [15, 10, 12]. Prior evaluations reveal that LLMs score poorly on social intelligence benchmarks, achieving only around 54% [28] on the SESI benchmark, suggesting underdeveloped strategies for handling nuanced interpersonal dynamics. In real-world applications, LLMs may analyze disputes effectively but struggle with impartial interventions, often overlooking ethical nuances or power imbalances [23, 16]

In this work, we treat compromise generation as a probe of model cognition. Given paired accounts describing the same type of place with opposing perspectives (e.g., safe vs. unsafe or welcoming vs. excluding) [4], the model should generate one or more compromises that should not favor either side. This controlled setting allows us to study how prompting strategies influence neutrality. In particular, we test whether iterative feedback, guided by an empathy-oriented similarity model, can reduce asymmetry in compromise generation. Particularly, we investigate the following research questions: **RQ1:** Have LLMs learned heuristics for empathic neutrality during pre-training, or do they exhibit biases toward asymmetric compromise generation when balancing divergent viewpoints? **RQ2:** Are LLM failures in gauging empathic neutrality intrinsic, or do they diminish when the model receives explicit neutrality guidance and examples?

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

To address these questions, we adapt the dataset by Chen et al. [4]. Each data point pairs two viewpoints on a public space: a safe/welcoming description (view_A) and a less safe/exclusionary description (view_B), each with a place description, reasons, and suggested improvements (see section A). From this corpus, we randomly selected 1200 welcome/excluded pairs and 1200 safe/unsafe pairs for our study. A suitable compromise must integrate both suggestions while remaining empathically neutral, which we define as minimizing the difference in empathic similarity between the compromise and each viewpoint ($|score_A - score_B| \to 0$), where $score_{A/B}$ denotes the empathic similarity between the generated compromise and view_A/view_B. This structure enables controlled evaluation of how models respond to divergent perspectives, as each compromise task draws on authentic, perspective-driven narratives rather than generic oppositional statements. We design prompt-based compromise generation methods and evaluate their effectiveness in a 50-participant study. We find that feedback-guided prompting produces more balanced compromises than single-prompt or basic CoT approaches. We also show that smaller open-source models can be aligned to approximate this ability; details of the alignment procedure are deferred to the appendix. Together, these results establish compromise generation as a compact, cognitively meaningful probe of how models integrate opposing cues and exhibit neutrality in social reasoning.

2 Compromise Generation using Empathy-Informed LLM

Compromise generation provides a concrete testbed for studying how LLMs integrate contrasting viewpoints into a single response. Our approach grounds this process in human-collected, contrasting viewpoints of safe/unsafe and welcome/excluded, and it examines the use of empathic similarity as a criterion for producing balanced, neutral compromises.

To generate compromises for each pair of contrasting views (including place description, reasons, and improvement suggestions), we employ Claude 3 Opus [1] via Alpaca-style prompting template [24], as single prompts (see Appendix C) often favor one. We generate multiple compromises per pair to enhance diversity and model generalization, exploring three strategies:

(i) *CoT*: Initially, we applied a basic CoT approach [6, 27], assuming that step-by-step reasoning might help the model identify commonalities and merge suggestions. In practice, however, the generated compromises often leaned too heavily toward one view or expanded beyond scope, producing unbalanced outputs.

(ii) *CoT+LLM*: This approach was inspired by prior work on self-generated feedback in LLMs [17]. The model produced both a compromise and empathic similarity score between each viewpoint and generated compromise, aiming for high similarity with minimal discrepancy. While this method modestly improved balance, the self-assigned scores often diverged from human judgments, limiting their reliability for supervision.

(iii) *CoT+Feedback*: This approach (Figure1), introduced an external loop by incorporating a separately trained similarity model to provide empathic similarity scores between each viewpoint and the generated compromise. Unlike self-evaluation, these scores mimic human empathic judgments and serve as objective feedback to guide iterative refinement. The method consistently improved neutrality and perceived quality, making CoT+Feedback the most effective strategy.

To compute similarity scores in *CoT+Feedback*, we adopted the empathic similarity framework from Shen [21]. For training the model, we used the EMPATHICSTORIES dataset (2,000 annotated pairs) augmented with 1,000 additional pairs from Chen et al. [4] of contrasting viewpoints (total: 3,000 pairs; 75/5/20 train/validation/test split). We used e5-large [26] for sentence representations. During inference, the model estimated empathic similarity between each viewpoint and the generated compromise (Refer to Appendix B) for more details).

3 Evaluation of Candidate Compromises for Empathic Neutrality

We evaluated the generated compromises both empirically and using human subjects.

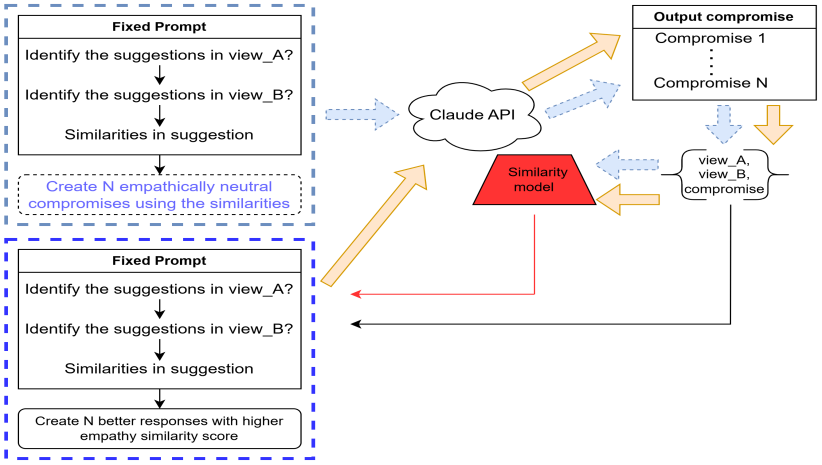

Figure 1: Iterative prompt engineering for compromise generation using similarity model guidance. Arrows show initial generation and refinement iterations.

Table 1: Distribution of candidate compromises across prompting strategies

Type	**Single Prompt**	**CoT**	**CoT+ LLM**	**CoT+ FB**
Welcome	0.75%	24.70%	28.45%	46.10%
Safe	0.83%	22.09%	26.65%	50.43%

Table 2: Preference distribution (%) from user study.

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

3.1 Empirical Evaluation of Compromise Methods

For each strategy (plus a single-prompt baseline), we generate four compromises per view pair (16 total). Out of these 16 compromises per view pair, we select the top four as candidate compromises based on neutrality between the empathic similarity scores of view_A, view_B and compromise (i.e., we want $|score_A - score_B| \to 0$). Table 1 shows the distribution: CoT+FB yields the most neutral candidates (46-50%), while single prompts yield the least (0.75%-0.83%).

3.2 Human Evaluation of Compromise Methods

We asked humans to rate the compromises generated by the three methods, Single Prompt, *CoT*, and *CoT+Feedback* for a sample of 100 viewpoint pairs. 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 identify with. The four other statements were generated compromises: one by Single Prompt, one by *CoT*, and two by the *CoT+Feedback* method. Each of participant rated the compromises for five pairs of statements.

The results of the study (Table 2) reveal a clear hierarchy in the effectiveness of the methods evaluated. As expected, the opposing statement, view_B, was never selected as first preference; it was selected as second preference only 1% of the time. The *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 two *CoT+Feedback* generated compromises (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.

Both the evaluation based on computed empathic similarity and the human evaluation indicate that:

- The LLM did not exhibit the ability to generate empathically neutral pairs based on its pretraining.
- The LLM performed poorly when asked to predict how empathically neutral a compromise text is with respect to two viewpoints.
- The LLM can generate more empathically neutral text when given better, external estimates of neutrality in a training scenario that incorporates iterative feedback.

For qualitative evaluation of section 2, Please see Appendix D.

3.3 A Closer Evaluation of Empathic Neutrality

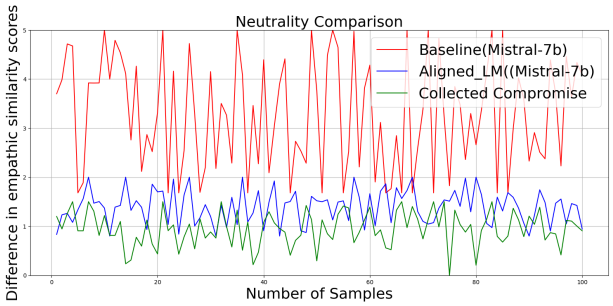

Figure 2: Score difference (lower = more neutral): Baseline model shows higher differences (less neutral), while iterative prompt engineering achieves lowest differences (most neutral). Task-based loss alignment significantly improves neutrality over baseline.

We compared the differences between the empathy similarity scores for view_A and view_B, towards a generated compromise. As shown in Figure 2, our candidate compromises have the lowest difference, indicating better empathic neutrality. The base LLM (mistral-7B [11]) had the worse performance, indicating LLM failures in gauging empathic neutrality are not intrinsic but arise from pretraining biases favoring academic over social intelligence. In order to reduce the gap, we found aligning the LLM towards these candidate compromises generate the desired ability (Due to space constraints, full algorithmic details and ablations are in the Appendix (summarized in Section F); qualitative examples are in Section E; neutrality gains are in Figure 2.).

4 Discussion and Conclusions

From our different prompt-based strategies, we observed that Claude Opus tended to generate compromises that were not empathically neutral and instead were biased towards one of the viewpoints. This answers the first part of RQ1: it did not learn to generate empathiccally neutral text and instead exhibited biases. With our *CoT+Feedback* approach, where feedback is iteratively given using an empathic similarity model to estimate neutrality, the LLM did learn to generate more empathically neutral compromises. When the *CoT+LLM* model evaluated how empathically neutral the generated compromise was, the estimates were much less effective in producing empathically neutral compromises, and so we infer that the LLM did not estimate empathic neutrality very well based on pre-training. However, employing outside evaluation of empathic similarity was effective and the model had the capacity to learn to generate more empathically neutral texts. We also observed that the score difference for the two views was smallest for *CoT+Feedback* and close when an LLM is aligned using an loss capturing empathic neutrality. These experiments answered RQ2, in that when the models received explicit neutrality guidance and examples, they were better able to generate empathically neutral text.

References

[1] AI Anthropic. Introducing claude, 2023.

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

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

[4] Francine Chen, Scott Carter, Tatiana Lau, Nayeli Suseth Bravo, Sumanta Bhattacharyya, Kate Sieck, and Charlene C Wu. Empathy prediction from diverse perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8959–8974, 2025.

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

[6] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 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, 2024.

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

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

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

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

[11] 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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[12] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 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, 2019.

[13] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

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

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

[16] Zilin Ma, Nathan Zhao, Linn Bieske, Blake Bullwinkel, Yanyi Zhang, Ziqing Luo, Siyao Li, Gekai Liao, Boxiang Wang, Jinglun Gao, et al. Using large language models for humanitarian frontline negotiation: Opportunities and considerations. *arXiv preprint arXiv:2405.20195*, 2024.

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

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

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

[21] Jocelyn Shen. *Modeling empathic similarity in personal narratives*. PhD thesis, Massachusetts Institute of Technology, 2023.

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

[23] Jinzhe Tan, Hannes Westermann, Nikhil Reddy Pottanigari, Jaromír Šavelka, Sébastien Meeùs, Mia Godet, and Karim Benyekhlef. Robots in the middle: Evaluating llms in dispute resolution. In *Legal Knowledge and Information Systems*, pages 168–179. IOS Press, 2024.

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

[25] 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. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

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

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

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

A Data sample

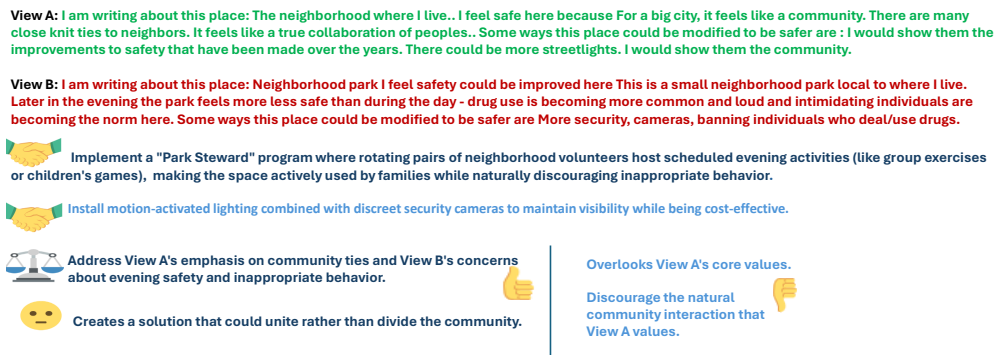

Figure 3: Example of a data point. View_A and view_B are collected from human participants, compromises ( ) are synthetically generated using prompt engineering that satisfy the criteria of balanced view ( ) and empathic neutrality ( ).

B Prompts for other CoT based strategy

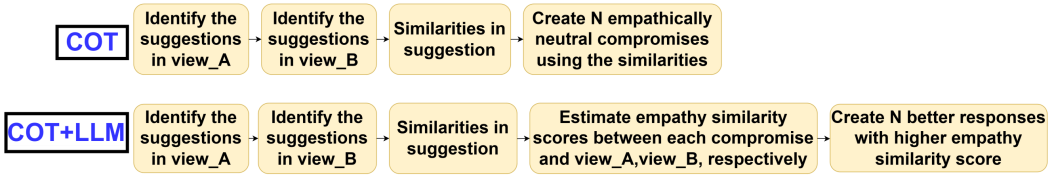

Figure 4: Prompt engineering strategies for collected compromise generation (a) basic COT approach (b) CoT+LLM

C Single prompt

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]

D Qualitative analysis of 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.

Results

Single Prompt

- 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.

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]

E 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)+Finetune: 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-7b): 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-7b)+Finetune: 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-7b: 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.]

F Aligning Foundation Models for Empathically Neutral Compromise Generation

Generating desired compromises requires a complex pipeline with prompt engineering, filtering, and proprietary models. To enable efficient, scalable inference, we align open source language models to directly produce empathically neutral compromises. Alignment typically uses reinforcement learning like PPO [18] with rewards (explicit, e.g., via GPT-4/Claude ratings [13], or implicit from preferences) or simplified methods like DPO [20], which optimizes likelihood ratios but is limited to pairwise data and may reduce absolute likelihood of optimal responses. Noise-contrastive estimation (NCE) [9]-based alignment [5] addresses this by optimizing absolute likelihood, and our initial NCE

application favors neutral compromises over fine-tuning alone. However, model likelihoods often misalign with task-specific metrics (e.g., BLEU [19], ROUGE [14]) [3]. To bridge this gap, we introduce a joint objective that optimizes both likelihood and task-based losses, yielding outputs that better reflect desirable compromise characteristics.

F.1 Task-loss based alignment

Due to the discrepancy between likelihood and task metric (in our case ROUGE score), we jointly optimize both. This encourages the model to generate compromises that are associated with higher likelihood and exhibit stronger alignment with candidate compromise in terms of ROUGE score, as formalized in Equation 1:

$$\begin{aligned} \mathcal{L}_{\text{tbl}} = \max(0, s_{target} - s_{hypo} + \\ |target_{rouge} - hypo_{rouge}| \cdot w\text{margin}) \end{aligned} \tag{1}$$

In this formulation, $target_{rouge}$ denotes the ROUGE score of the candidate compromise with respect to itself (used as reference), while $hypo_{rouge}$ is the ROUGE score of the fine-tuned base model generated hypothesis with respect to the corresponding candidate compromise. The term w_{margin} (weight margin) modulates the influence of ROUGE difference during training.