LLM Alignment for Linguistic Empathy: A Reinforcement Model Learning from Symbolic Feedback (RLSF)

Anonymous ACL submission

Abstract

Research on conversational AI and large language models (LLMs) has highlighte need to find ways to align LLMs with human's expectations towards empathetic dialogue. This paper presents a new approach towards alignment via reinforcement learning from symbolic feedback (RLSF) drawing on the theory of linguistic empathy. Instead of relying on relying on humans' evaluations of an LLM's empathy and using Reinforcement learning from human feedback (RLHF) to align LLMs, we propose a method that in which a synthetic agent evaluates a LLM's generated responses based on their consistency with theoretically grounded synthetic properties of empathy (i.e., active versus passive, pronouns, tense, and the imperative form). To realize such RLSF, we first trained a reward model that learns an observable reward signal incorporating all four syntactic rules, followed by policy learning to align an LLM for better empathy (e.g. using proximal policy optimization, PPO). Our results indicate the potential of the RLSF approach in rendering more empathetic LLMs.

1 Introduction

Large Language Models (LLMs) have become the backbone of today's conversational AI. Scholars raise concerns that LLMs might not align with human expectation towards empathetic responses. Broadly speaking empathy requires LLMs to understand and share the feeling of others (Cuadra et al., 2024). Several techniques have been used to achieve better alignment with humans' need for empathy. Reinforcement learning with human feedback (RLHF) is the most popular method currently used and studied (Ziegler et al., 2020; Kaufmann et al., 2023).

Aligning LLMs towards empathy is challenging given the lack of high quality labelled data causing inconsistencies and biases in the alignment (Kaufmann et al., 2023). Empathy is perceived and interpreted differently depending on the conversational context and the emotional state of the human (). To overcome this challenge, we turn to linguistic theories of empathy that focuses on the syntactic properties of language structure that imbue the artificial conversational agent with an empathetic capacity without changing the underlying meaning and logic of the language (Kuno and Kaburaki, 1977). The basic premise of linguistic empathy is that empathy relates to lexical choices of the speaker that manifest in the points of view expressed by speakers and understood by listeners in a conversation (Kuno and Kaburaki, 1977; Kann, 2017; Kann et al., 2023). Such syntactic properties can be formally modeled as symbolic rules. Building upon these theoretical foundations, we propose a new approach for LLM alignment using reinforcement learning (RL). We first train a reward model that learns a hidden state of sentence's empathy emerging from the combination of four syntactic rules of linguistic empathy applied to response data (answers) in a conversational setting. This learned reward model, representing the symbolic feedback (SF) of an "artificial" expert, is used to fine-tune an LLM using deep RL model (e.g. using policy optimization (PPO). We refer to this approach as reinforcement learning from symbolic feedback (RLSF). In this paper, we present a pilot implementation of our RLSF architecture and evaluate it.

2 Related Work

Our research relates to the following literature streams: LLM alignment and linguistic empathy.

2.1 LLM alignment

In the setting of conversational AI, seminal papers and surveys highlight the need for empathy alignment. Existing work uses prompt engineering and learning-oriented approaches to tackle this need.

The existing learning-oriented approaches can be generally classified in 1) reinforcement learning (RL) and supervised learning (SL) methods. Alignment with RLHF is currently the most widely used approach of LLM alignment, and thus, the focus in this current paper. The most popular RLHF models for LLM focus on alignment for the purpose of a helpful, honest, and harmless (**?**Wu et al., 2021). The basic idea of RLHF is a combination of inverse reinforcement learning (Arora and Doshi, 2021) and reinforcement learning, where an agent learns behavior from rewarded interactions with its environment(Ziegler et al., 2020; Kaufmann et al., 2023; Bai et al., 2022).

While there has been significant progress in the algorithms for reward learning and policy learning, RLHF bears significant challenges. The most significant challenges relate to the sampling of human feedback data itself due to human biases (e.g. the evaluation of harmful versus helpful is subjective and biased based on a human's context and past experience) as well as 2) sampling challenges (e.g. lack of external validity, inter coder reliability challenges, etc.), and 3) lack of theoretical grounding of psychological and linguistic concepts like empathy. Further, our review of the RLHF literature shows there is no focused effort that tackles the specific nature of empathy and how it is expressed through language. Our research aims to these challenges with a focus on a theoretically grounded approach.

2.2 Empathy prediction

Empathy recognition is a nascent field of research in computational linguistics. Scholars have started to train multimodal language models to detect empathy using text, speech, and also video data (Hasan et al., 2023, 2024b,a). One of the shortcomings of such empathy recognition models is that they rely on human labels for empathy rather than theoretical grounded measurement techniques of empathy used in psychometrics (Kann, 2017). In the context of LLM alignment and empathetic language generation, empathy has become an important concept used for LLM evaluation (Cuadra et al., 2024). However, empathy is typically conceptualized in a very generic way (e.g. helpful versus harmful) (Bai et al., 2022). One of the major shortcomings in prior work, is that scholars ignore that empathy is a distinct property of language structure that is displayed through language syntax and grammar (Kuno and Kaburaki, 1977; Kann et al., 2023; Kann, 2017).

3 RLSF: Theory and technical implementation

3.1 Theoretical foundations

We draw upon the linguistic theory of empathy in language structure (Kuno and Kaburaki, 1977; Kuroshima and Iwata, 2016). Linguistic theory of empathy explains that empathy can be displayed through syntax: subtle changes in syntax structure significantly impact empathy display, even if the underlying logic of the sentence remains the same (Kuno and Kaburaki, 1977; Kuroshima and Iwata, 2016). For example, structuring a sentence in active tense rather than passive tense ("John hit Mary" versus "Mary was hit by John") changes the perspective of the receiver because it shifts the "camera angle" (Kuno and Kaburaki, 1977). Table 4 in appendix A presents four synthetic rules, derived from linguistic empathy theory. Each of the four rules related 1) active versus passive voice, 2) pronouns, 3) tense, and 4) imperative, explain how a distinct property of syntax structure change the "camera angle" in the conversation, and thus, display greater degree of empathy.

3.2 Model architecture

Figure 1 presents our architecture. It combines 1) symbolic feedback data based on linguistic empathy, 2) reward modeling, and 3) reinforcement policy learning using PPO.

3.2.1 Symbolic agent and empathy rules

We start by constructing a dataset of questions with empathetic and non-empathetic answers. We prompted[1] gpt-4o to create tenant-law legal questions in the Chicago area considering several topics such as security deposits, leases, agreements, et cetera. The dataset consists of 2000 questions, an accepted answer, and a rejected one, covering our four theoretical elements of empathy in language (i.e., active versus passive, present tense versus past tense, the imperative form versus *please do [verb]*, and personal pronouns for showing transitivity and reflexivity of verbs).

3.2.2 Reward modeling

We trained a reward model using our joint preference dataset as shown in table 7 in appendix A. We fine-tuned the DeBERTa V3 model (He et al., 2020) on the accepted-rejected pairs in our dataset. Our choice DeBERTa is that it outperforms RoBERTa

[1]The prompts used are shown in table 5 appendix A

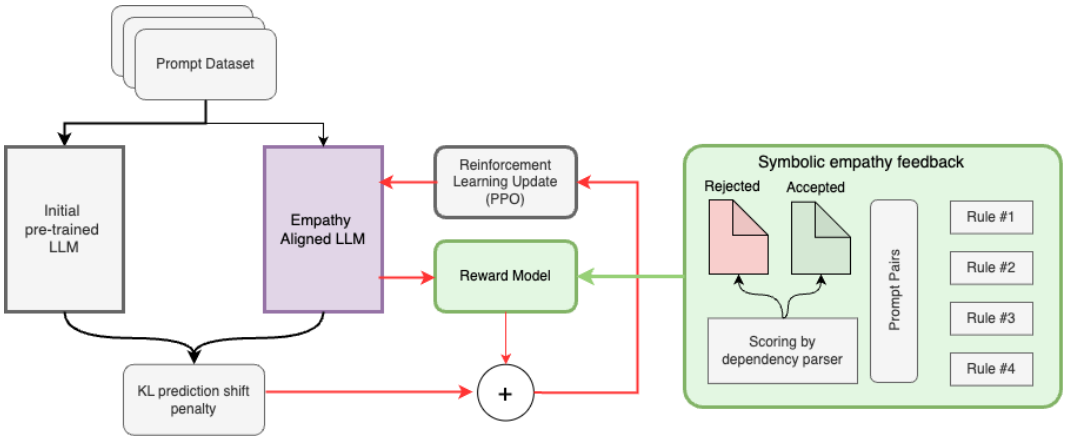

Figure 1: Proposed RLSF architecture

on several NLU tasks as it improves the BERT and RoBERTa models using disentangled attention and enhanced mask decoder. We fine-tuned it for 500 steps with a training batch size of 32 and a validation batch size of 16. We report a training loss of *0.0001*, a validation loss of *0.0106*, and an accuracy of *0.995*.

Metric (Macro)	Reward Model
Accuracy	94.00
Precision	94.11
Recall	94.00
F1	93.99

Table 1: These results show the metrics of evaluating the reward model.

We evaluated the reward model on a test dataset that includes 400 question-answer pairs. Each linguistic rule has 100 question-answer pairs (50 questions with an answer with a ground truth label of 1 and 50 questions with a ground truth label of 0). Table 1 shows the macro F1 score.

3.2.3 Reinforcement learning (RL) using PPO

We used the Huggingface trl (von Werra et al., 2020), whose PPO trainer aligns a generated response with a given query based on rewards from our fine-tuned reward model. In each step, a batch of prompts is sampled from the dataset. These prompts are then used to generate responses from the supervised fine-tuning model. The reward model evaluates these responses to compute rewards, which are subsequently used to optimize the supervised fine-tuning model through the PPO algorithm. We use the default generation parameters (i.e., generation kwards)[2]. We used the quantized 4bit Llama 3 model from unsloth[3]. The model was trained for 84 iterations for 5 hours on two GPUs, NVIDIA L40S[4]. We showed the model the data 5 times (5 epochs).

3.3 Model evaluation

3.3.1 Training metrics

Overall, we achieved stable training results. The average entropy score achieved was 22.42 (std: 0.29), and the kl distance was 0.15 (sd:0.153), the kl coefficient was 0.198 (sd: 0.0017). Overall, this suggests that the training was robust and the model did not diverge unexpectedly. The average response length token was 53.39 (sd: 12.15626), and the average computed reward reward increased over time, and stayed positive. Figure 2 in appendix A shows the KL divergence.

3.3.2 Evaluation of responses after prompting

We evaluated the performance of our alignment using RLSF by comparing the aligned LLM and the baseline LLM (4bit Llama 3 model) on 100 questions. First, we compared the similarity of the responses to those 100 questions using BLEU score, ROUGE score, Levenshtein Distance, and Semantic Similarity.

Table 2 presents the mean scores of four different metrics used to evaluate the performance of our aligned LLM against the baseline LLM. The BLEU mean score indicates a moderate level of precision in producing text similar to the reference answers

[2]min_length=-1, top_k=0.0, top_p=1.0, do_sample=True, max_new_tokens=32, pad_token_id=tokenizer.eos_token_id

[3]https://huggingface.co/unsloth/llama-3-8b-Instruct-bnb-4bit

[4]hyperparameters: learning rate=$1.41e-5$, batch size=512, minibatch size=16

Metric	score
BLEU (Reiter, 2018)	23.80
ROUGE (Zhang et al., 2024; Lin, 2004)	0.40
Levenshtein Distance (Yujian and Bo, 2007)	55.67
Semantic Similarity (Reimers and Gurevych, 2019)	0.85

Table 2: Mean scores of the similarity metrics

while the ROUGE mean score suggests a low-to-moderate overlap in word sequences, meaning that the aligned model can capture content relevance. A significant number of character changes is needed, according to the Levenshtein distance, to transform the model's output into the baseline text, indicating some differences in the exact wording. Finally, the semantic similarity score indicates a high degree of similarity between the output of the two models.

We also scored our LLMs, the baseline and the aligned model using our dependency parser (see Table 3). The mean values suggest that despite the short training time and small sample size of our pilot, the aligned model is more empathetic. More training, evaluation, and advanced statistical analysis is required to make statistical claims of mean differences.

Rule	aligned	baseline
Rule 1: Active	0.7 (1.0)	0.67 (1.0)
Rule 2: Pronouns	0.04 (0.0)	0.02 (0.0)
Rule 3: Tense	0.92 (1.0)	0.92 (1.0)
Rule 4: Imperative	1.0 (1.0)	1.0 (1.0)

Table 3: Mean and (median) scores of the evaluations using the dependency parsing.

3.3.3 Comparison with "human" baseline

We also compared our aligned and our unaligned model with a human baseline. We used an open dataset published by () [sabine] add yifan reference here, in which the researchers had handcrafted a bot using the symbolic rules introduced in **??**

4 Discussion and future work

Our proposed architecture for aligning LLMs for empathy using symbolic feedback replaces human feedback with symbolic feedback produced by a synthetic agent using syntactic rules. Our major contribution of this paper is the 1) conceptual clarification of linguistic empathy, 2) the design and implementation of a symbolic feedback agent for linguistic empathy 3) the training of a reward model for linguistic empathy. and 4) the implementation and evaluation of RLSF alignment architecture. It is worth noting that despite the lack of data accessible for baseline evaluation, we also compared our alignment model with a "human expert" baseline, a handcrafted bot that was designed by a team of experts guided by the theory of empathy in language display (Brunswicker et al., 2024).

5 Limitations

There are several limitations that set the stage for future work. study does not fully capture that empathy is a multidimensional concept. We only focus on linguistic empathy, and limit ourselves to synthetic rules. We only represent four syntactic rules but ignore linguistic elements of empathy that are expressed through rhetoric. Future work may incorporate rhetoric rules that complement syntactic rules. However, dependency parsing may be insufficient for creating and/or evaluating such symbolic rules.. In addition, our pilot implementation bears technical limitations. First, we only generated a small dataset for our reward modeling that was generated using a combination of dependency parsing and prompt engineering. In the future, we plan to advance our reward modeling significantly by leveraging existing datasets created for emotion recognition and parsing it using our symbolic rules .Thus, alternatives to PPO might be explored. Finally, our evaluation is also limited. For example, our RLSF alignment method should be compared with prompt engineering methods, prompting the LLM to use the four rules individually and jointly. While our results indicate that RLSF with empathy rules may bear significant benefits, more data, training and evaluation is needed. Thus, future work may invest in additional data generation for symbolic feedback in order to produce not only training but also test data. Further, we plan to compare PPO with other optimization methods as well as DPO.

References

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken

conversations. *Computer Speech & Language*, 50:40–61.

Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. ArXiv:2204.05862 [cs].

Sabine Brunswicker, Yifan Zhang, Christopher Rashidian, and Dan W. Linna Jr. 2024. *The Impact of Empathy Display in Language of Conversational AI: A Controlled Experiment with a Legal Chatbo*.

Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure? In *4th Conference on Conversational User Interfaces*, pages 1–13, Glasgow United Kingdom. ACM.

Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024a. LLM-GEm: Large Language Model-Guided Prediction of People's Empathy Levels towards Newspaper Article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231, St. Julian's, Malta. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, Susannah Soon, and Shafin Rahman. 2023. Curtin OCAI at WASSA 2023 Empathy, Emotion and Personality Shared Task: Demographic-Aware Prediction Using Multiple Transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 536–541, Toronto, Canada. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Aneesh Krishna, Shafin Rahman, and Tom Gedeon. 2024b. Thesis Proposal: Detecting Empathy Using Multimodal Language Model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 338–349, St. Julian's, Malta. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ilona Herlin and Laura Visapää. 2016. Dimensions of empathy in relation to language. *Nordic Journal of Linguistics*, 39(2):135–157. Publisher: Cambridge University Press.

Trevor Kann. 2017. *Measuring Linguistic Empathy: An Experimental Approach to Connecting Linguistic and Social Psychological Notions of Empathy*. Ph.D. thesis, University of California, Los Angelos.

Trevor Kann, Steven Berman, Michael S. Cohen, Emily Goldknopf, Merve Gülser, Gennady Erlikhman, Kristi Trinh, Olga T. Yokoyama, and Eran Zaidel. 2023. Linguistic Empathy: Behavioral measures, neurophysiological correlates, and correlation with Psychological Empathy. *Neuropsychologia*, 191:108650.

Timo Kaufmann, Sarah Ball, Jacob Beck, Eyke Hüllermeier, and Krauke Kreuter. 2023. On the Challenges and Practices of Reinforcement Learning from Real Human Feedback. Turin, IT.

Susumu Kuno and Etsuko Kaburaki. 1977. Empathy and Syntax. *Linguistic Inquiry*, 8(4):627–672. Publisher: The MIT Press.

Satomi Kuroshima and Natsuho Iwata. 2016. On Displaying Empathy: Dilemma, Category, and Experience. *Research on Language and Social Interaction*, 49(2):92–110.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Grant Packard, Sarah G. Moore, and Brent McFerran. 2018. (I'm) Happy to Help (You): The Impact of Personal Pronoun Use in Customer–Firm Interactions. *Journal of Marketing Research*, 55(4):541–555.

Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv:1908.10084 [cs].

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 27(4):582–584.

Li Yujian and Liu Bo. 2007. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ming Zhang, Chengzhang Li, Meilin Wan, Xuejun Zhang, and Qingwei Zhao. 2024. ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics. *Expert Systems with Applications*, 237:121364.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. ArXiv:1909.08593 [cs, stat].

A Appendix

No.	Description	No Empathy	Empathy
1	Active vs passive (Herlin and Visapää, 2016)	Third person singular. (*"This problem can be solved"*).	First person plural; second person if necessary (e.g. "We can solve this together", or "You can solve this problem"
2	Pronouns (Packard et al., 2018)	No personal pronouns (no reflexivity or transitivity of words), e.g. *The deposit*.	Personal pronouns (combined with active tense & transitivity and reflexivity of verbs) e.g. *Your deposit*.
3	Tense ((Kuroshima and Iwata, 2016))	Past tense	Present tense
4	Imperative (Alam et al., 2018; Cox and Ooi, 2022; Pelau et al., 2021)	Imperative statement with first word being verb, e.g. *"Send a letter to your landlord"*	Empathetic Imperative (Please + do + Infinitive)., e.g. *"Please do send a letter to your landlord.*

Table 4: Syntactic rules for linguistic empathy; based on (Kuno and Kaburaki, 1977; Brunswicker et al., 2024; Kann, 2017)

empathy feature	prompt
active vs. passive	You are a tenant-law lawyer: can you generate questions about [topic] in Chicago area? Avoid generating yes/no questions. For each of these questions: Generate two answers: one in active voice and another in passive voice.
present vs. past	You are a tenant-law lawyer: can you generate questions about [topic] in Chicago area? Avoid generating yes/no questions in a format of a table. For each of these questions: Generate two answers: one in present tense and another in past tense
imperative	You are a tenant-law lawyer: can you generate questions about [topic] in Chicago area? Provide two answers for each question. One with the imperative form and one starting with the word "please", followed by "do", then whatever verb after that.
pronouns	You are a tenant-law lawyer: can you generate questions about [topic] in Chicago area? Avoid generating yes/no questions. Include the output in a format of a table. For each of these questions: Generate two answers: one that includes the pronouns (your, our) and another without these pronouns.

Table 5: Prompts used to generate tenant-law legal questions considering several topics such as security deposits,leases, agreements, general rights and responsibilities of the tenant and the landlord, rents, payments, disputes, legal actions, evictions, moving out, et cetera.

question	accepted	rejected
Human: How can tenants document the condition of their rental property before moving out in Chicago?	Assistant: You should take detailed photos and videos of the entire property, noting any existing damages or wear.	Assistant: Detailed photos and videos of the entire property should be taken, noting any existing damages or wear.
Human: How can a landlord address tenant violations of lease terms in Chicago?	Assistant: A landlord can address tenant violations by issuing a notice of violation, setting a deadline for corrective action, and pursuing eviction if the issues are not resolved.	Assistant: Addressing tenant violations of lease terms includes issuing a notice of violation, setting a deadline for corrective action, and pursuing eviction if issues are not resolved by the landlord.

Table 6: Examples of the preference data following the HH-RLHF (**?**) dataset style. The questions are used in the PPO pipeline.

accepted	rejected
Human: How can tenants document the condition of their rental property before moving out in Chicago? Assistant: You should take detailed photos and videos of the entire property, noting any existing damages or wear.	Human: How can tenants document the condition of their rental property before moving out in Chicago? Assistant: Detailed photos and videos of the entire property should be taken, noting any existing damages or wear.
Human: How can a landlord address tenant violations of lease terms in Chicago? Assistant: A landlord can address tenant violations by issuing a notice of violation, setting a deadline for corrective action, and pursuing eviction if the issues are not resolved.	Human: How can a landlord address tenant violations of lease terms in Chicago? Assistant: Addressing tenant violations of lease terms includes issuing a notice of violation, setting a deadline for corrective action, and pursuing eviction if issues are not resolved by the landlord.

Table 7: Examples of the joined preference data following the HH-RLHF (**?**) dataset. This dataset is used to train the DeBERTa reward model.

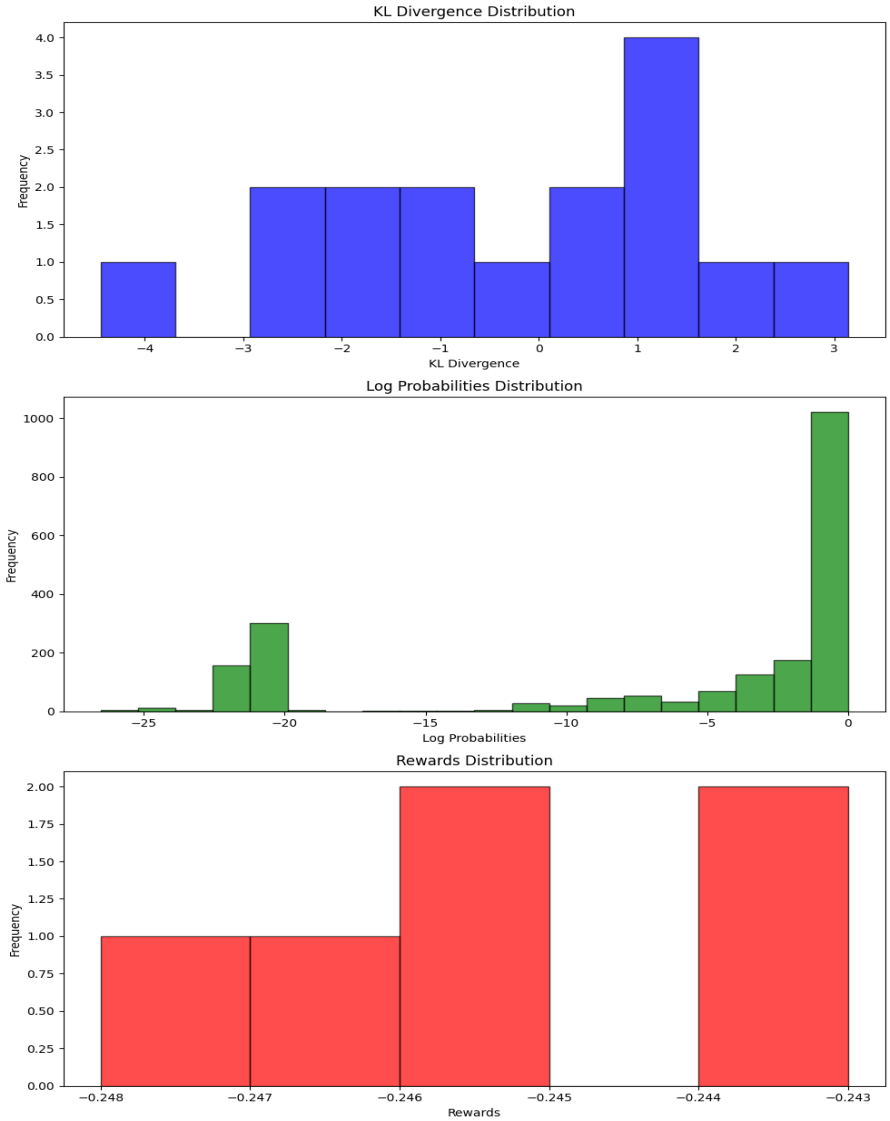

Figure 2: These plots show the distributions of KL divergence, log probabilities, and rewards