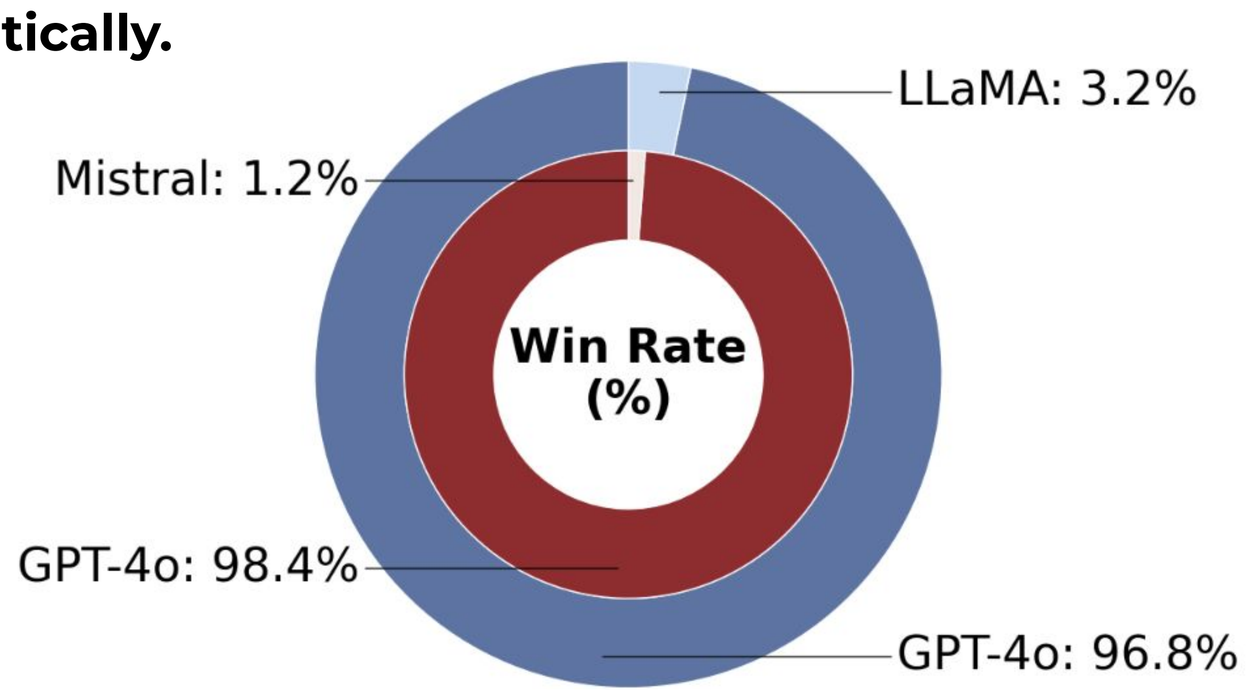
Distilling Empathy from Large Language Models

Problem Statement

Empathy plays a crucial role in positive human interactions and communication [1]. Large Language Models (LLMs) have shown proficiency in understanding emotions and responding in empathetic, supportive ways, but they are **expensive and require a lot of computing power.** Small Language Models (SLMs) are much cheaper and easier to deploy, but **often struggle to respond empathetically.**



GPT-4o vs. Base LLaMA-3.1-8B & Mistral-7B-v0.3 in generating empathetic responses as judged by Gemini; Graphic created by finalist using matplotlib, 2026

Goal: Develop an empathy distillation framework that systematically transfers the empathetic capabilities of LLMs into SLMs [3]

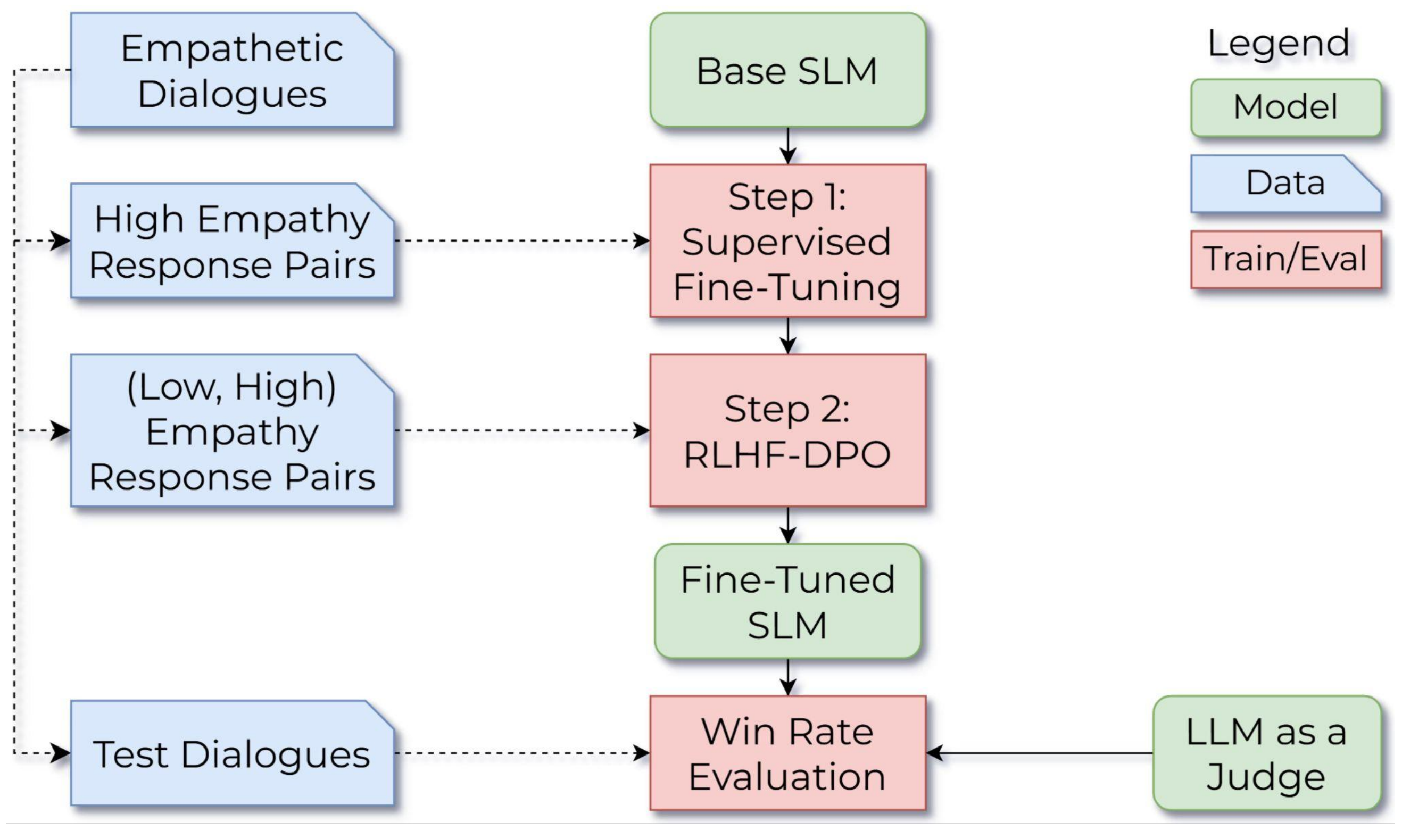
Research Questions: 1. How can we methodically create datasets for empathy distillation from LLMs? 2. How can we efficiently utilize empathy distillation datasets to fine-tune SLMs?

Our Approach

One comprehensive approach to distilling empathy from LLMs into SLMs

- Two-step fine-tuning
  - Supervised Fine-Tuning (SFT) first with high empathy responses
  - Reinforcement Learning from Human Feedback (RLHF) through Direct Preference Optimization (DPO) [2] second with (low, high) empathy response pairs
- Three different methods to distill empathy from LLMs
  - Direct empathy distillation
  - Targeted empathy improvement over human responses
  - Targeted empathy improvement over LLM initial responses
- Four prompting strategies with significant improvement over distilling empathy through direct prompting

Two-Step Fine-Tuning



Two-Step Fine-Tuning Process; Graphic created by finalist with draw.io, 2026

Three Empathy Distillation Methods

Method 1: Direct Empathy Distillation

- Use LLMs' high-empathy responses [4] to fine-tune SLMs

Method 2: Targeted Empathy Improvement over Human Responses

- Four prompt families based on the naive prompt that refine human replies along the cognitive, affective, and compassionate dimensions of empathy

Method 3: Targeted Empathy Improvement over LLM Initial Responses

- Same prompts as Method 2
- Instead of improving over human initial responses, Method 3 improves over LLM initial responses, eliminating the need for human examples.
- Because we do not have human empathy scores to partition the SFT and RLHF datasets, we adopt the same SFT and RLHF split as Method 2.

Four Prompting Strategies

Naive Prompt
Below is a response to a given speaker utterance in a given context. Generate a new improved empathetic response, using on average 28 words and a maximum of 97 words, that is of higher empathetic quality and also retains the original meaning, intention, and emotion of the original response.

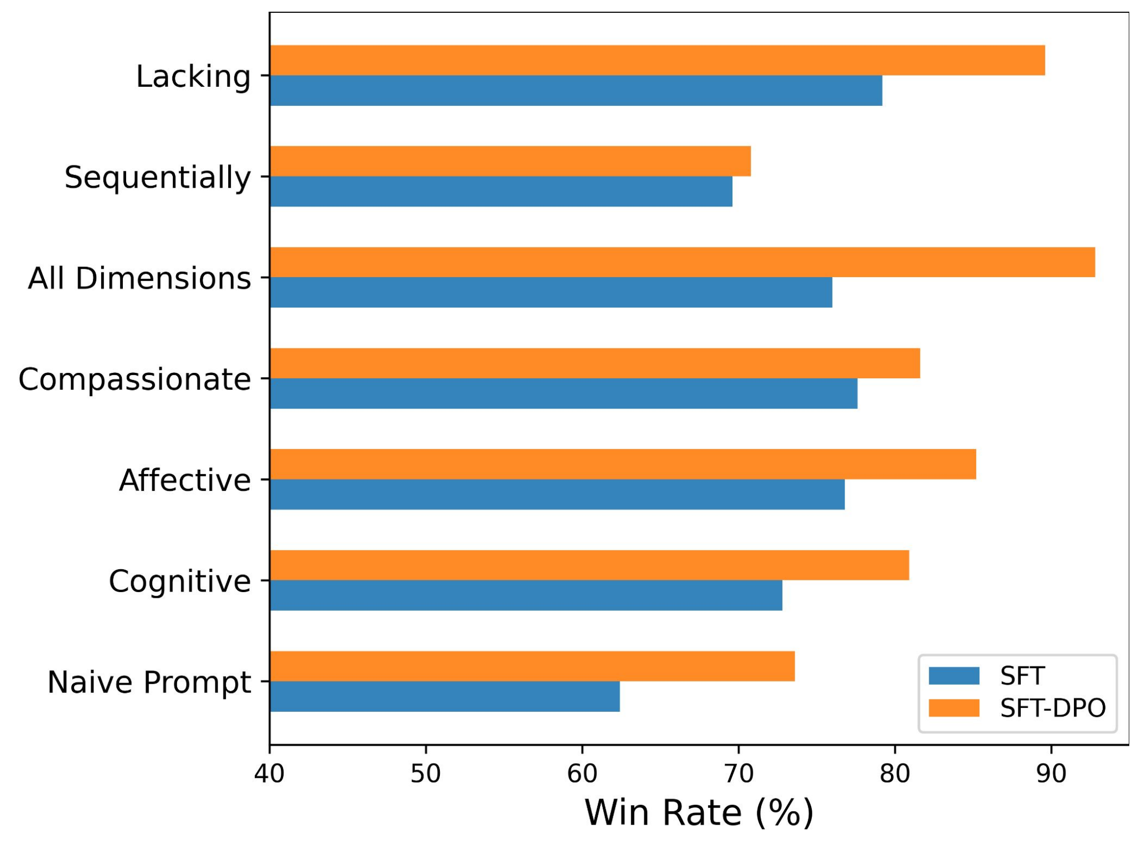
Prompt 1: Improve Along One Dimension of Empathy
{Naive Prompt}
{Strategy} Your higher quality response should be improved specifically along the [cognitive, affective, compassionate] dimension of empathy.
{Definition of [cognitive, affective, compassionate] dimension of empathy}

Prompt 2: Improve All Three Dimensions of Empathy
{Naive Prompt}
{Strategy} Your higher quality response should be improved along the three dimensions of empathy: cognitive, affective, and compassionate empathy.
{Definitions of cognitive, affective, and compassionate dimensions of empathy}

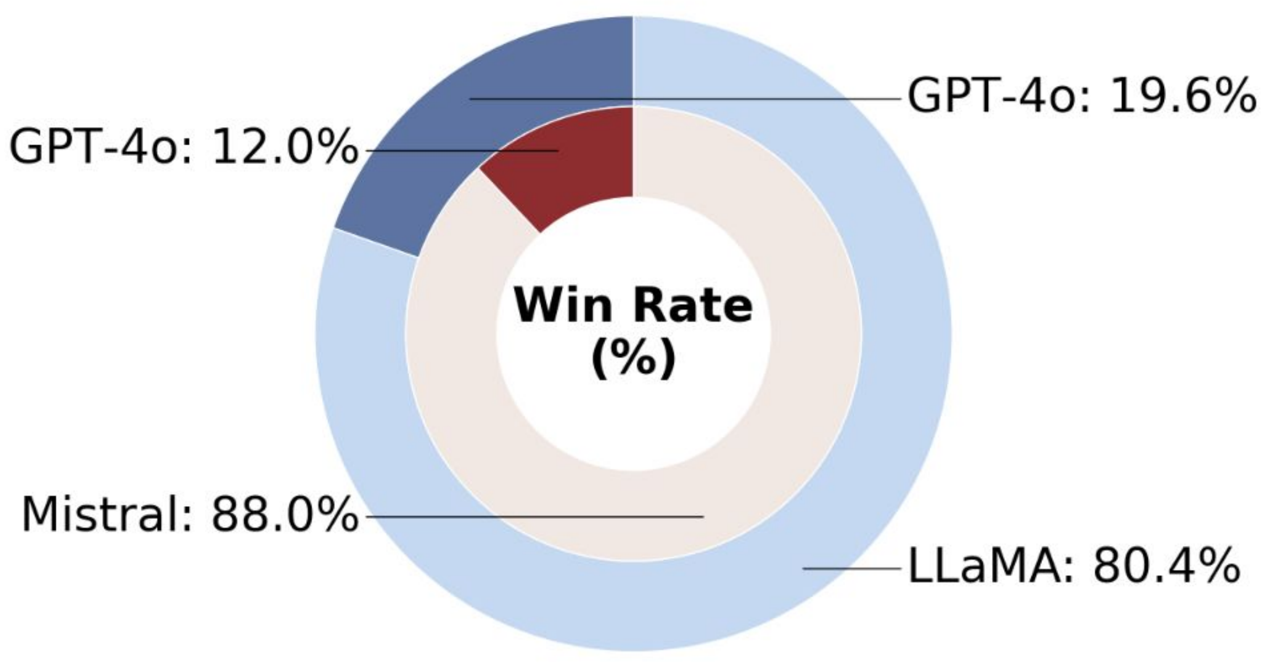
Prompt 3: Improve Three Dimensions Sequentially
{Naive Prompt}
{Strategy} Your higher quality response should be improved specifically along the 1. cognitive, 2. affective, and 3. compassionate dimensions of empathy.
{Definition of 1. cognitive, 2. affective, and 3. compassionate dimensions of empathy}

Prompt 4: Identify the Lacking Dimension
{Naive Prompt}
{Strategy} In the process of generating a higher quality empathetic response, you should identify the dimension of empathy (cognitive, affective, and compassionate dimensions) that the original response lacks most of, and specifically improve along the lines of the dimension you identified.
{Definitions of cognitive, affective, and compassionate dimensions of empathy}

Experimental Results



Win-rates of Fine-Tuned LLaMA-3.1-8B vs. Base LLaMA-3.1-8B using different targeted empathy improvement prompts and fine-tuning methods; Graphic created by finalist using matplotlib, 2025 and published in SIGDIAL 2025



GPT-4o vs. Fine-Tuned LLaMA-3.1-8B & Mistral-7B-v0.3 in generating empathetic responses as judged by Gemini; Graphic created by finalist using matplotlib, 2026

Takeaways

- SLMs fine-tuned through the two-step process with distillation datasets enhanced by the targeted empathy improvement prompts significantly outperform the base SLMs at generating empathetic responses with a win rate of 90+%.
- The best fine-tuned SLMs even outperform the state-of-the-art LLMs in generating empathetic responses.

References

1. Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of Personality and Social Psychology, 44(1):113–126.
2. Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems (NeurIPS).
3. Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. LLM pruning and distillation in practice: The Minitron approach. Arxiv preprint, abs/2408.11796.
4. Anuradha Welivita and Pearl Pu. 2024. Are large language models more empathetic than humans? Arxiv preprint, abs/2406.05063.

Henry Xie