Cognitive Feedback: Decoding Human Feedback from Cognitive Signals

Yuto Harada
The University of Tokyo, Japan
harada-yuto@g.ecc.u-tokyo.ac.jp

Yohei Oseki
The University of Tokyo, Japan
oseki@g.ecc.u-tokyo.ac.jp

Abstract

Alignment from human feedback has played a crucial role in enhancing the performance of large language models. However, conventional approaches typically require creating large amounts of explicit preference labels, which is costly, time-consuming, and demands sustained human attention. In this work, we propose Cognitive Feedback, a framework that infers preferences from electroencephalography (EEG) signals recorded while annotators simply read text, eliminating the need for explicit labeling. To our knowledge, this is the first empirical investigation of EEG-based feedback as an alternative to conventional human annotations for aligning language models. Experiments on controlled sentiment generation show that CPO achieves performance comparable to explicit human feedback, suggesting that brain-signal-derived preferences can provide a viable, lower-burden pathway for language model alignment.

1 Introduction

Human alignment for large language models (LLMs) is crucial for generating safe and preference-aligned outputs. Previous work has shown that this process helps LLMs better follow human instructions and mitigate harmful behaviors (Ouyang et al., 2022). A traditional post-training approach involves supervised fine-tuning (SFT) on a pretrained LLM, followed by reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020). Direct Preference Optimization (DPO) (Rafailov et al., 2024) is an alternative to RLHF that skips the reward model and offers more stable training. Many state-of-the-art models, such as OpenAI's o-series, continue to adopt the SFT + DPO paradigm (Guan et al., 2024), demonstrating that it remains an effective strategy. However, creating the preference labels necessary for DPO and related preference optimization methods remains labor-intensive. Tasks such as selecting and training annotators, establishing trust, and coordinating large-scale annotation efforts incur substantial costs (Stiennon et al., 2020; Casper et al., 2023a).

To address these challenges, Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023) leverages LLM-generated synthetic feedback to substitute for explicit human feedback. This approach offers lower costs, easier large-scale data collection, and strong scalability compared to traditional human-driven methods (Wang et al., 2022; Madaan et al., 2024; Bai et al., 2022). However, several drawbacks remain. Depending on the task, humans may disagree with AI-generated judgments (Perez et al., 2022; Casper et al., 2023b; Lee et al., 2023), indicating that synthetic feedback may fail to capture genuine human intentions. Moreover, there is a bootstrapping issue: ensuring the model that produces feedback is itself properly aligned is non-trivial (Casper et al., 2023a), theoretically undermining AI feedback as a complete solution to alignment. Finally, while AI-generated feedback can reduce cost, it does so at the expense of direct human involvement, raising concerns about whether such signals faithfully reflect nuanced human values. The question of which feedback signals, or combinations of such signals, most effectively align LLMs with human goals remains open (Casper et al., 2023a).

In this work, we propose Cognitive Feedback, a framework for obtaining preference information directly from human brain activity. Specifically, we investigate whether preference signals extracted from electroencephalography (EEG) can be integrated into preference optimization methods such as DPO. If feasible, this approach could offer a more direct and potentially less cognitively demanding means of capturing individual responses than conventional annotation pipelines, as participants only need to read the presented text without providing explicit ratings. We focus on the con-

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 209–219

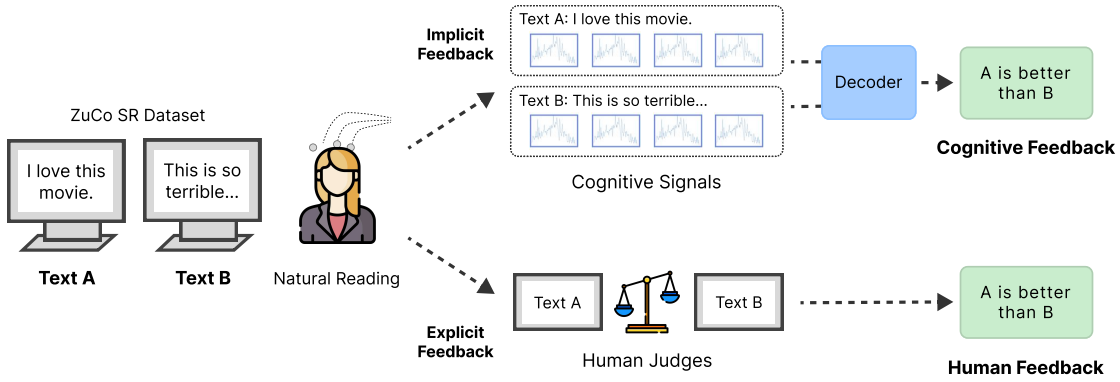

Figure 1: A diagram depicting Cognitive Feedback (top) vs. Human Feedback (bottom). By decoding human preferences from biosignals, it enables obtaining implicit human feedback without explicit annotation.

trolled sentiment generation task. This task is well suited to our study for two reasons: (1) it serves as a foundational benchmark for preference optimization, as it was one of the benchmark tasks originally employed in DPO or other derivative methods (Rafailov et al., 2024; Zeng et al., 2024; Amini et al., 2024), and (2) previous works have demonstrated that EEG is effective at capturing emotional responses in NLP (Wang and Zhang, 2025). To operationalize this idea, we introduce Cognitive Preference Optimization (CPO), a method that estimates preference information from EEG data collected while participants read text. By relying on implicit cognitive feedback instead of explicit human feedback, CPO aims to significantly reduce the need for manual annotation (Figure 1). Alongside falling costs and growing accessibility in mobile EEG, recent large-scale decoding results show clear data–performance scaling, reinforcing the practical path for EEG-based alignment (Sato et al., 2024).

In our experiments, we compare two forms of feedback: standard human feedback requiring explicit labeling, and implicit feedback inferred from EEG. Our results show that the CPO-trained model not only produces more positive outputs than a baseline model but also achieves performance comparable to conventional human feedback settings. These findings highlight the potential for EEG-based feedback signals to serve as a novel approach for LLM alignment.

We summarize our main contributions:

1) We propose Cognitive Feedback, a framework that replaces explicit annotations with implicit feedback decoded from EEG collected during natural reading.

2) We instantiate this framework with a DPO-based method that uses EEG-decoded preferences (CPO), empirically demonstrating the feasibility of using EEG signals to guide preference optimization on a controlled sentiment generation task.

3) We compare CPO with conventional human feedback or AI feedback, illustrating that EEG-derived feedback can effectively align language models while potentially reducing the burden of manual annotation.

2 Related Works

2.1 Cognitively Inspired Natural Language Processing

Previous studies have shown that incorporating human physiological signals can boost performance in a variety of natural language processing (NLP) tasks. In particular, eye-tracking data has been employed to improve part-of-speech tagging (Barrett et al., 2016), text simplification (Klerke et al., 2016; Higasa et al., 2024), dependency parsing (Strzyz et al., 2019), sentiment analysis (Barrett et al., 2018), named entity recognition (Hollenstein and Zhang, 2019), relation classification (Hollenstein et al., 2019; McGuire and Tomuro, 2021), text readability (González-Garduño and Søgaard, 2017; Hollenstein et al., 2022), and sarcasm detection (Mishra et al., 2016a,b, 2017). Across these diverse tasks, leveraging eye-tracking data has consistently led to notable gains in model performance.

Compared to eye-tracking, relatively few works have explored EEG signals for NLP. Nevertheless, several studies have established the effectiveness of EEG in tasks such as named entity recognition, relation extraction, and emotion classification (Hollenstein et al., 2019; Ren and Xiong, 2021). In

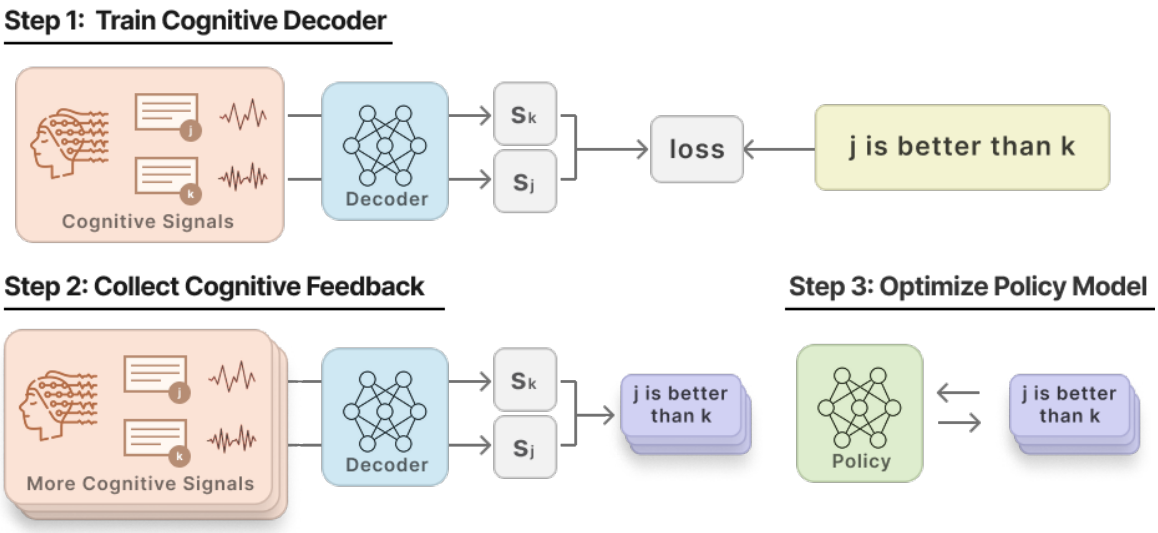

Figure 2: Schematic diagram of Cognitive Preference Optimisation. In Step 1, a decoder is trained using a small set of cognitive signals (e.g., EEG) paired with explicit human feedback; in Step 2, the trained decoder is used to infer preferences from a larger set of cognitive signals without manual labelling.

addition, Muttenthaler et al., 2020 regularized attention mechanisms with EEG data to improve performance on relation extraction, and Wang and Zhang, 2025 demonstrated that EEG can be a valuable modality for emotion detection. Most of these earlier approaches relied on encoder-only architectures, which cannot be directly applied to the decoder-only models now prevalent in NLP. Because architectural modifications are typically required, it is difficult to leverage existing pretrained models in these methods.

More recently, researchers have begun exploring how physiological signals can be integrated into post-training workflows for modern large language models (LLMs). For instance, Kiegeland et al., 2024a incorporated eye-tracking feedback into Direct Policy Optimization (DPO), while Lopez-Cardona et al., 2024 built a reward model by applying the synthetic gaze generation method proposed by Khurana et al., 2023 to create a large-scale dataset of artificially generated gaze data. Additionally, Kiegeland et al., 2024b applied eye-tracking to supervise a cognitive modeling step via supervised fine-tuning (SFT). Our work is the first to examine whether EEG data can be utilized for post-training alignment in modern LLMs.

2.2 Aligning Large Language Models with Human Feedback

Recent large language models (LLMs), such as GPT-4 (OpenAI et al., 2024), Llama 3 (Grattafiori et al., 2024), Claude 3 (Anthropic., 2024), and Gemini (Team et al., 2024), have demonstrated impressive capabilities across a wide range of tasks. These models are typically pretrained on massive datasets and then undergo post-training to better follow human instructions. One of the most common approaches for human alignment is Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020), which generally comprises three main steps: (1) collecting human feedback, (2) training a reward model (RM) based on that feedback, and (3) optimizing the LLM via a reinforcement learning algorithm such as Proximal Policy Optimization (PPO) (Schulman et al., 2017). Since RLHF was first introduced, numerous improvements have been proposed, such as fine-grained reward systems (Bai et al., 2022; Wu et al., 2023b; Dong et al., 2023; Wang et al., 2023, 2024) and alternative RL methods that replace the original PPO module (Wu et al., 2023a). Beyond RLHF, (Rafailov et al., 2024) proposed Direct Preference Optimization (DPO), an offline RL approach that optimizes language models directly on preference data without training a separate reward model. DPO has been shown to provide training stability and match the efficacy of RLHF. Notably, even state-of-the-art models continue to adopt these methods, often combining supervised fine-tuning (SFT) with DPO to achieve strong performance on a variety of tasks.

However, a primary limitation of RLHF lies in the difficulty of data collection, which encompasses issues such as evaluator misalignment, supervisory challenges, and variable feedback quality (Casper

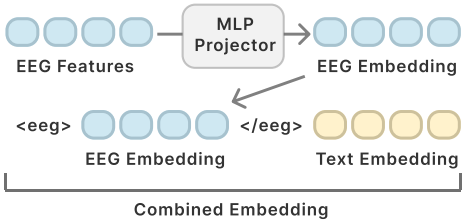

Figure 3: Concat text and EEG embedding with randomly initialised special tokens.

et al., 2023a). To address these problems, recent studies have shifted focus toward AI-generated feedback. For instance, Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022) and its variants (Lee et al., 2023; Zhu et al., 2024; Cui et al., 2023; Li et al., 2024; Yang et al., 2023) leverage synthetic feedback from LLMs, greatly reducing labeling costs and improving scalability. That said, these approaches do not fully resolve the drawbacks of RLHF. Depending on the task, humans often disagree with AI-generated judgments (Perez et al., 2022; Casper et al., 2023b; Lee et al., 2023). The disagreement rate varies widely—for example, Perez et al. (2022), Casper et al. (2023b), and Lee et al. (2023) report figures of up to 10%, 46%, and 22%, respectively, in different experiments. Furthermore, it remains unclear which forms of feedback signals, or which combinations thereof, most effectively align LLMs with human goals (Casper et al., 2023a), indicating a need for continued exploration.

3 Cognitive Preference Optimization

As outlined in Section 2, cognitive signals on their own can be noisy; however, they serve to enrich NLP embeddings by providing more detailed information. We adopt this paradigm for AI Feedback: cognitive signals function as an implicit form of human feedback, capturing user preferences with minimal burden on the annotators, while reinforcing the input information used in AI feedback. In so doing, we attempt a novel feedback approach that alleviates the limitations of both human and AI feedback. Figure 2 is an overview of the proposed method.

Step 1: Training Cognitive Decoder Let $X = (x_1, x_2, \ldots, x_T)$ be the sequence of combined feature vectors for a text of length T. Following previous work (Lopez-Cardona et al., 2024), each x_t is formed by concatenating the EEG feature vector

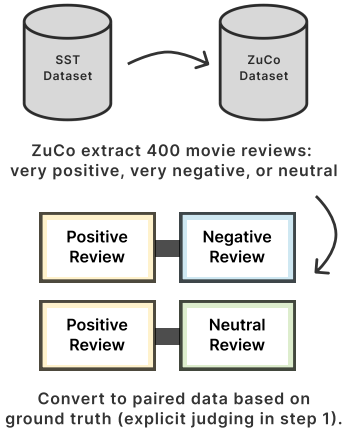

Figure 4: Overview of the preparation of the preference pair dataset used in our experiments.

$e_t \in R^m$ (recorded when reading the t-th token) with its text embedding $h_t \in R^{m'}$, i.e. $x_t = [e_t; h_t]$. This approach has been shown to yield robust representations (Figure 3). We then define a Cognitive decoder $s_\phi(X) \in R$, where ϕ denotes its trainable parameters. For training, suppose we have N pairs $\{(X^{(i)}_{\text{chosen}}, X^{(i)}_{\text{rejected}})\}_{i=1}^{N}$. We want the decoder to assign a higher score to $X^{(i)}_{\text{chosen}}$ than to $X^{(i)}_{\text{rejected}}$. To achieve this, we minimize:

$$\mathcal{L}(\phi) = \sum_{i=1}^{N} \log\Big(1 + \exp\big(-[s_\phi(X^{(i)}_{\text{chosen}}) - s_\phi(X^{(i)}_{\text{rejected}})]\big)\Big) \tag{1}$$

which encourages $s_\phi(X^{(i)}_{\text{chosen}})$ to be larger than $s_\phi(X^{(i)}_{\text{rejected}})$.

Step 2: Collecting Cognitive Feedback Next, we use the trained Cognitive Decoder to collect cognitive feedback. Although preference data were required as supervision in Step 1, Step 2 only requires EEG signals. Specifically, given two candidate texts, we compute their scores with Cognitive Decoder. We designate the text with the higher score as `chosen` and the one with the lower score as `rejected`, thus creating a pair of texts with corresponding preference information. This approach reduces the need for explicit human annotation.

Step 3: DPO with Cognitive Feedback Finally, we use the cognitive feedback gathered in Step 2 as preference data to optimize a language model via Direct Policy Optimization (DPO). DPO maximizes the likelihood that preferred outputs are selected over less-preferred ones, relative to a reference model, and it does so without requiring a separate reward model. Formally, given a model

π_θ and a reference model π_{ref}, DPO minimizes the following loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log \sigma\Big(\beta\Big[\log \frac{\pi_\theta(\boldsymbol{y}^{(i)}_{\text{chosen}} \mid \boldsymbol{x}^{(i)})}{\pi_{\text{ref}}(\boldsymbol{y}^{(i)}_{\text{chosen}} \mid \boldsymbol{x}^{(i)})} - \log \frac{\pi_\theta(\boldsymbol{y}^{(i)}_{\text{rejected}} \mid \boldsymbol{x}^{(i)})}{\pi_{\text{ref}}(\boldsymbol{y}^{(i)}_{\text{rejected}} \mid \boldsymbol{x}^{(i)})}\Big]\Big) \tag{2}$$

where β is a temperature-like hyperparameter, $\boldsymbol{y}^{(i)}_{\text{chosen}}$ denotes the chosen output for the i-th text $\boldsymbol{x}^{(i)}$, and $\boldsymbol{y}^{(i)}_{\text{rejected}}$ is the rejected output. In this way, the model is optimized to align its generation with the preferences inferred from the EEG signals, effectively reducing the need for explicit human labels.

4 Experiments

In this section, we empirically evaluate the performance of our proposed method by examining three questions: (1) To what extent can we decode feedback from EEG signals? (2) Does the proposed method perform at a level comparable to conventional, explicit human feedback? (3) Does its performance scale with the size of the EEG datasets we use? Although no EEG dataset currently exists for the purpose of LLM preference optimization, if initial experiments demonstrate the method's effectiveness even under limited data conditions, this would provide motivation for creating larger, more realistic datasets. This work serves as a first step toward assessing whether cognitive signals can supplement or even replace traditional forms of human feedback.

4.1 Preference Pair Dataset Processing

In this work, we extract cognitive signals from an existing natural reading corpus and convert them into pairwise preference data (Figure 4). Notably, the participants' task did not involve reading pairs of texts for direct comparison; instead, they read single texts and attempted to infer their sentiment labels. This discrepancy between the participants' reading task and the NLP objectives is a disadvantageous setup that may complicate improvements in performance.

Dataset We use the Sentiment Reading (SR) dataset from the Zurich Cognitive Language Processing Corpus (ZuCo) (Hollenstein et al., 2018), which captures both eye-tracking and EEG data simultaneously. This makes ZuCo particularly suitable for NLP tasks requiring word-level EEG features. The SR subset comprises about 400 movie-review sentences, read by 12 participants. These sentences were drawn from the Stanford Sentiment Treebank (SST) (Socher et al., 2013), focusing on clearly positive, negative, and neutral sentences to ensure representative samples for each sentiment category. We extract EEG features at the word level based on Gaze Duration (GD), resulting in 840-dimensional vectors per word. ZuCo is currently the largest dataset that meets the requirements of our experiments.

Conversion to Pairwise Preference Data Because the SR dataset in ZuCo was not originally intended for reinforcement learning, we convert its single-sentence labels into pairwise preference data. The SR set contains 400 sentences labeled according to the Stanford Sentiment Treebank (SST): 140 positive, 137 neutral, and 123 negative. To avoid data leakage, we split these sentences into 10 folds while preserving their label distribution, and construct pairwise preferences based on the relations *positive* $>$ *neutral* and *positive* $>$ *negative*. Although we could theoretically create all possible pairs (e.g., each positive sentence paired with every neutral or negative one), we restrict each sentence to at most five pairs during training to mitigate overfitting due to repetitive examples. At test time, however, we generate as many pairs as possible. The EEG decoder is trained via 10-fold cross-validation, and from each test fold we obtain all qualifying pairs, yielding a total of 3,640 pairs used as cognitive feedback.

Human Feedback Collection Out of the 400 sentences in the SR dataset, 47 have five-level sentiment ratings provided by human annotators. Among these 47 sentences, the ground-truth distribution is 22 positive, 6 neutral, and 17 negative. Based on these labels, we create a total of 506 pairs. For each pair, we derive a preference signal from the five-level sentiment rating, which serves as human feedback. Because the number of human feedback samples is relatively small, we select the same 506 text pairs from the cognitive feedback set for direct comparison. This ensures that any difference in performance arises from the feedback source, rather than from inconsistencies in the underlying data.

	Input Type	Model			
		Llama-3-8B		Llama-3-8B-Instruct	
Baseline	Text	79.3 ± 0.6	diff (%)	79.1 ± 0.9	diff (%)
Cognitve Decoder	Text + EEG	**82.9 ± 1.1***	4.5	**81.6 ± 0.5***	3.2
	Text + Noise	75.4 ± 2.8	-4.9	77.4 ± 2.3	-2.2

Table 1: Cognitive Decoder Accuracy (%) for ZuCo SR dataset. Highest results are in bold; "diff" indicates rate of improvement and reports statistical significance.

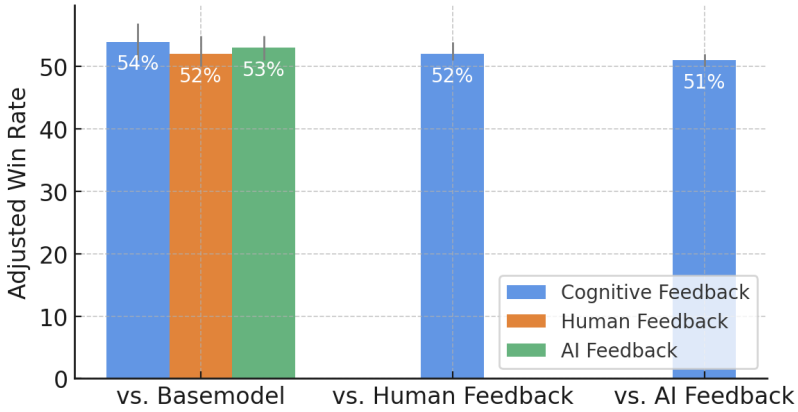

Figure 5: Performance of policies trained with Cognitive Feedback. Direct comparison with policies from other feedback types and indirectly through comparison with the base model.

4.2 Settings

Cognitive Decoder The Cognitive Decoder takes as input a sequence of embeddings derived from text and cognitive signals, producing a higher score for texts deemed more positive. Following the methodology of (Lopez-Cardona et al., 2024), we used the pretrained Llama-3-8B and Llama3-8B-Instruct (Grattafiori et al., 2024) models as decoders. However, rather than the standard classification head for next-word prediction, we replaced it with a regression head that outputs a scalar score (Touvron et al., 2023).

Policy Model The policy model is optimized to generate more positive movie reviews. We use GPT-2-large (774M parameters)[1] as our base model. We found that gpt-2-medium produced lower quality text, so we used a larger model. These findings are similar to those in (Rafailov et al., 2024). During training, we employ a common prompt, "`movie review: `", to encourage consistent outputs.

[1] https://huggingface.co/openai-community/gpt2-large

Tasks We empirically evaluate the performance of our proposed method using a single controlled sentiment generation task (Rafailov et al., 2024), for which we employ two types of prompts. The first prompt, referred to as the "SST Prefix Prompt," leverages the Stanford Sentiment Treebank dataset: we select 50 neutral sentences and 50 negative sentences (none of which overlap with the ZuCo SR dataset), and provide only the initial 10 words of each sentence (the prefix) to the policy model, which then generates the continuation. The second prompt, referred to as the "Training Condition Prompt," aligns more closely with our training conditions. In both cases, we allow up to 50 tokens to freely continue from each prompt.

Evaluations We conduct two types of evaluations on the texts generated for the tasks described above. The first is an llm-as-a-judge approach, where we use GPT-4o-2024-11-20 to select which model produces the more positive output. We evaluate the output with the following prompt: "`Which is the more positive movie review? Please write this down as (A) or (B). If you feel equally positive, answer (C).`" Based on these selections, we compute an adjusted win

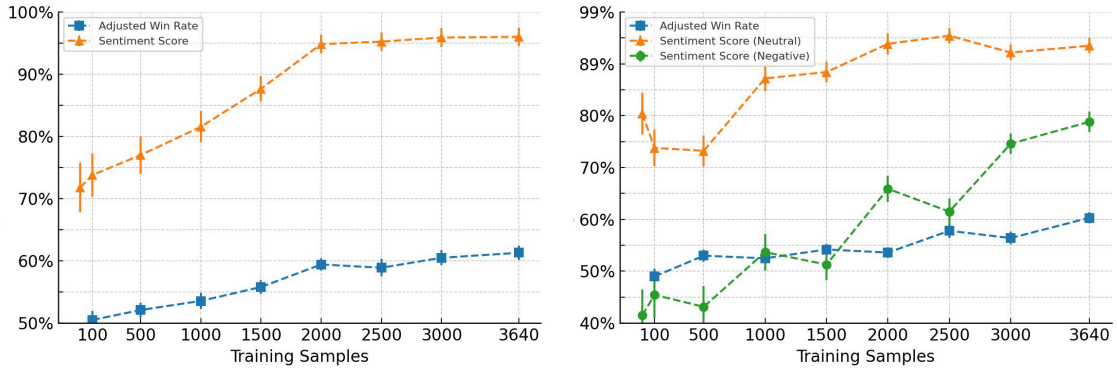

Figure 6: Correspondence between the number of training samples and performance. Results with training prompts (left) and with SST prefixes (right).

rate to assess each policy model. In addition to comparing a model trained with human feedback against one trained with cognitive feedback, we also compare each trained model with an untrained base model. The latter comparison indirectly evaluates the two trained models' performance. The second method employs a sentiment classification model to verify that the generated text is genuinely positive. We adopt a RoBERTa-large (Hartmann et al., 2023) fine-tuned on the IMDb dataset, which uses the probability score for the positive label in a binary classification to evaluate each generations.

4.3 Results

Decoding Feedback from EEG Signals We illustrate the performance of the EEG decoder in Table 1. In the Baseline setting, no EEG features are used as inputs; rather, the model predicts scores solely from text embeddings. In our experiments, this simple text-based output serves as the "AI Feedback." In contrast, the Cognitive Decoder takes both text embeddings and EEG features as its input representations. To verify the contribution of EEG data to the decoding task, we also experimented with random noise vectors that have the same dimensionality as the EEG embeddings. Our results indicate that combining text information with EEG features yields higher-accuracy feedback decoding, consistent with findings in prior research on cognitively inspired NLP. Moreover, the fact that random noise not only fails to improve performance but degrades it suggests that the EEG features indeed contain task-relevant information. We use the outputs decoded by the Cognitive Decoder, which we refer to as "Cognitive Feedback," to train the policy model. Meanwhile, the test outputs decoded under the Baseline setting are used as "AI Feedback".

Cognitive Feedback vs. Human Feedback vs. AI Feedback We show the performance of the policy trained with cognitive feedback, compared to those trained with human feedback and AI feedback, in Figure 5. The adjusted win rate is computed using the "llm-as-a-judge" approach and reflects the average score across two prompt types, evaluated over five trials. Note that human feedback is available for only 47 out of 400 sentences in the ZuCo SR dataset, representing only a portion of the entire dataset. For fairness in comparison, the data used for other feedback types is restricted to this same subset. Despite the smaller training set, the policy trained with cognitive feedback outputs more positive text than the base model, which does not undergo reinforcement learning, and achieves a higher win rate. Its performance is comparable to, or slightly surpasses, that of the other feedback types. One possible explanation is that the cognitive feedback approach, much like AI feedback, draws on text embeddings but further leverages EEG signals to augment these embeddings, thereby potentially providing a more powerful input representation.

Scaling to number of training samples. Figure 6 illustrates how performance changes as we increase the number of training data pairs. In both prompt types, performance consistently improves with a larger number of pairs. Before undergoing reinforcement learning, the base model generates positive outputs more than 70% of the time for training prompts and around 80% of the time for a neutral SST prefix. However, for a negative prefix, it produces proportionally more negative outputs, indicating relatively natural behavior. As training progresses, however, the model gradually shifts toward producing positive continuations, even for

SST Prefix (Negative):	
Original:	**Do we really need a 77-minute film to tell us** exactly why a romantic relationship between a 15-year-old boy and a 40-year-old woman doesn't work? - NEGATIVE
Base model:	**Do we really need a 77-minute film to tell us** what happened? A quick glance at this Wikipedia page gives a bit of information. - NEGATIVE
CPO:	**Do we really need a 77-minute film to tell us** everything that we need to know about this game? Absolutely! The best part of this movie is how much the players of this great team seem to get into their characters. - POSITIVE
Training Prompt	
Base model:	**movie review:** I was a little apprehensive. "Avengers: Age of Ultron" is a great film. There are some really great characters and moments, and the story is a nice blend of action, comedy, and drama. - POSITIVE
Base model:	**movie review:** I'm still not sure how to feel about the new video game from the creators of Batman: Arkham Origins. While it has all the trappings of a video game I'd rather not play — no cutscenes, no stories. - NEGATIVE
CPO:	**movie review:** ""A dazzling and stirring gem that will continue to inspire generations of filmgoers."" – James Bobin, National Board of Review - POSITIVE
CPO:	**movie review:** A smart, witty, and highly entertaining film about a family's remarkable journey of faith and growth. - POSITIVE

Table 2: Example of a model trained by the proposed method and the generated text of the base model. Each sentence was labelled using the sentiment classification model used to evaluate the model.

negative prefixes. Alongside the observed improvement in win rate, it is clear that the model increasingly favors affirmative or positive statements.

5 Discussion

Examples of text generated by the proposed method are presented in Table 2. The CPO model shown in this table was trained with the maximum number of available preference pairs, representing its best-performing configuration in our experiments. For prompts with the "SST prefix" type, even when the initial text begins with a clearly negative statement, the CPO model often changes the tone partway through the continuation and shifts the overall sentiment toward a more affirmative or optimistic direction. As a result, the generated sentences sometimes receive sentiment labels that differ from those assigned to the original prompt. For the "training prompt" type, the base model generally produces continuations that are emotionally neutral or slightly positive, but these outputs can still be classified as neutral or negative by the sentiment classifier. In contrast, the CPO model consistently produces continuations in this setting that are classified as positive, indicating that the EEG-derived feedback effectively steers generation toward more favorable sentiment across both prompt types.

6 Conclusion

In this paper, we proposed Cognitive Preference Optimization (CPO), a novel framework for aligning large language models (LLMs) with human preferences inferred from electroencephalography (EEG) signals. By training a cognitive decoder to extract pairwise preferences from a natural reading corpus, we introduced a method that reduces reliance on explicitly labeled data. Our results suggest that EEG-derived feedback can successfully guide policy optimization for sentiment generation, producing outputs that match or even rival models trained with conventional human feedback. The proposed method can use the scalability of traditional AI feedback while obtaining human feedback in the form of readings that are less burdensome for the operator. Future experiments in a more realistic setting will require the construction of a large dataset of cognitive signals for the purpose of reinforcement learning of LLMs.

Limitations

The experiments in this study focus only on a controlled sentiment generation task, so it is not yet clear whether EEG-derived feedback is effective for more complex or open-ended tasks. The current method also estimates preferences only in a pairwise comparison setting, without exploring scalar or multi-dimensional feedback that could provide richer training signals. We report performance with GPT-2-large for generation and Llama-3-8B for EEG decoding, chosen given the modest corpus size; effectiveness is not guaranteed when larger baseline models are used.

Ethical Considerations

This work uses only publicly available and properly licensed datasets that permit research use. All datasets were used in accordance with their intended research purposes. AI tools were used solely to assist in writing training and analysis scripts.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24H00087 and JST PRESTO Grant Number JPMJPR21C2.

References

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*.

Anthropic. 2024. Claude 3: Introducing the next generation of claude.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023a. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023b. Explore, establish, exploit: Red teaming language models from scratch. *Preprint*, arXiv:2306.09442.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Taichi Higasa, Keitaro Tanaka, Qi Feng, and Shigeo Morishima. 2024. Keep eyes on the sentence: An interactive sentence simplification system for english learners based on eye tracking and large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing nlp with cognitive language processing signals. *Preprint*, arXiv:1904.02682.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel Kiegeland, David Robert Reich, Ryan Cotterell, Lena Ann Jäger, and Ethan Wilcox. 2024a. The pupil becomes the master: Eye-tracking feedback for tuning llms. In *ICML 2024 Workshop on LLMs and Cognition*.

Samuel Kiegeland, Ethan Wilcox, Afra Amini, David Robert Reich, and Ryan Cotterell. 2024b. Reverse-engineering the reader. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9367–9389, Miami, Florida, USA. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv e-prints*, pages arXiv–2309.

Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, and 1 others. 2024. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. *arXiv preprint arXiv:2403.08309*.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2024. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models. *arXiv preprint arXiv:2410.01532*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Erik McGuire and Noriko Tomuro. 2021. Relation classification with cognitive attention supervision. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 222–232, Online. Association for Computational Linguistics.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016a. Predicting readers' sarcasm understandability by modeling gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016b. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Berlin, Germany. Association for Computational Linguistics.

Lukas Muttenthaler, Nora Hollenstein, and Maria Barrett. 2020. Human brain activity for machine attention. *Preprint*, arXiv:2006.05113.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

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.

Yuqi Ren and Deyi Xiong. 2021. CogAlign: Learning to align textual neural representations to cognitive language processing signals. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3758–3769, Online. Association for Computational Linguistics.

Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. 2024. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *Preprint*, arXiv:2407.07595.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Towards making a dependency parser see. 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 1500–1506, Hong Kong, China. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jing Wang and Ci Zhang. 2025. Cross-modality fusion with eeg and text for enhanced emotion detection in english writing. *Frontiers in Neurorobotics*, 18:1529880.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and 1 others. 2023. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. 2023a. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment. *arXiv preprint arXiv:2310.00212*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrastive distillation for language model alignment. *arXiv preprint arXiv:2307.12950*.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.