NEUROPROMPT: EEG-CONDITIONED EMOTION-AWARE DIALOGUE GENERATION WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although large language models can generate fluent dialogue, emotional response generation often relies on explicit emotion cues or manual labels that are rarely available in realistic interaction. Since EEG signals have been widely studied as a physiological source for emotion recognition, they can serve as an implicit affect cue for language models. To address the lack of paired EEG-dialogue corpora, we propose a staged framework for EEG-conditioned dialogue generation. This framework learns cross-subject emotion EEG representations, maps them into the LLM embedding space, and uses the projected representation to guide emotion-adaptive response generation. Under a cross-subject protocol, blind GPT-5.4 evaluation on 1,598 paired dialogue samples shows that EEG-Conditioned LLM achieves a 31.93% win rate and 2.23 average rank, outperforming Text-Only Baseline (22.44%/2.45) and Unconditioned LLM (11.67%/3.17) while approaching Gold-Label Oracle (33.96%/2.14); a 500-sample human questionnaire shows the same trend. We further conduct an online real-time EEG user study in which each participant's EEG is fed to the model during live interaction, and anonymized responses are rated for how well they match the participant's current emotional state. In this online study, EEG-Conditioned LLM also achieves the best human ratings and average rank among the compared models. Taken together, the offline and online results support EEG as a practical source of implicit emotion cues for response adaptation and suggest that brain-derived signals can improve dialogue generation beyond text-only supervision.

1 INTRODUCTION

Large language models (LLMs) can now generate fluent and helpful dialogue, yet adapting responses to a user's current emotional state remains challenging in realistic interaction Rashkin et al. (2019); Wang et al. (2022). Existing approaches rely on explicit emotion labels or instructions Liu et al. (2023b), or infer emotion from dialogue text Wang et al. (2022). Explicit labels are rarely available in natural conversation, while text-only inference can miss feelings that users do not verbalize.

Electroencephalography (EEG) offers a physiological signal that can be measured during interaction and has been widely studied for emotion recognition Wang et al. (2014); Zheng & Lu (2015). However, most work focuses on classification, and recent EEG+LLM studies Kim et al. (2024); Jiang et al. (2025); Mishra et al. (2025) do not directly address its use as an implicit control signal for dialogue. This raises our central question: can EEG help LLMs generate responses whose style better matches the user's emotional state?

Several challenges remain. EEG and dialogue are highly heterogeneous modalities, and there is currently no large-scale corpus of synchronized EEG-dialogue pairs that would support end-to-end brain-to-response training. Affective EEG patterns also vary across individuals Zheng & Lu (2016), requiring generalization to unseen users. The key problem is to convert cross-subject EEG affect information into an LLM-compatible conditioning signal.

We propose a staged EEG-conditioned framework that learns cross-subject EEG emotion representations, maps them into the LLM's emotion-control-token embedding space, and uses the projected signal as a soft prompt. Text-side supervision and subsequent pair refinement train the generator un-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

der weakly paired supervision. At inference, the model receives dialogue context and EEG without explicit emotion labels.

On 1,598 pair-suitable test samples, blind GPT-5.4 evaluation gives the EEG-Conditioned LLM a 31.93% win rate and a 2.23 average rank, outperforming practical text-only baselines and approaching the Gold-Label Oracle. A 500-sample human questionnaire and a preliminary real-time user study show consistent benefits.

Our contributions are as follows.

- We study dialogue generation conditioned on EEG-derived emotion, treating EEG as an implicit control signal without requiring explicit emotion labels at inference time.
- We propose a staged framework under weakly paired supervision. It combines cross-subject EEG representation learning, lightweight projector mapping, and EEG-conditioned soft-prompt generation to connect raw EEG with dialogue generation.
- We show through blind LLM-judge evaluation, a human questionnaire, qualitative counterfactual examples, and ablations that EEG can serve as a practical control signal for affect-level response adaptation.
- We present a preliminary real-time online user study beyond the offline dataset setting. Participants freely compose utterances during live EEG-conditioned interaction, which remains rare in this line of work and provides user-side evidence under more open interaction conditions.

2 Related Work

2.1 EEG Emotion Recognition

EEG is a common physiological signal for emotion recognition because of its high temporal resolution and relative objectivity Wang et al. (2014). Datasets such as the SEED series Zheng & Lu (2015); Zheng et al. (2018); Liu et al. (2021); Jiang et al. (2024), DEAP Koelstra et al. (2011), and MAHNOB-HCI Soleymani et al. (2011) have made this a standard supervised benchmark. Prior methods include feature-based classifiers Duan et al. (2013); Wang et al. (2014), recurrent and compact convolutional models Alhagry et al. (2017); Lawhern et al. (2018), graph or Transformer architectures Li et al. (2022a); Song et al. (2018); Zhong et al. (2020); Ding et al. (2025), and self-supervised representation learning Li et al. (2022b). Because affective EEG varies across individuals, cross-subject generalization is also widely studied Li et al. (2018); Ma et al. (2019); Li et al. (2021). This literature mainly optimizes recognition accuracy, whereas our goal is to use EEG as a control signal for downstream response generation.

2.2 Emotional Dialogue and Multimodal LLMs

Emotion-aware dialogue generation has been studied on empathetic conversation benchmarks such as EmpatheticDialogues Rashkin et al. (2019). Prior systems guide generation using explicit target emotion labels Zhou et al. (2018), prompt-level guidance or demonstrations Qian et al. (2023), or predicted emotion signals from dialogue context or user utterances Wang et al. (2022). Recent extensions use multimodal emotion cues Liu et al. (2024) or vocal information Wu et al. (2025), but these settings remain closer to language and speech than EEG-based emotion control.

Multimodal LLMs similarly connect non-text signals to language models through lightweight bridges or frozen-backbone alignment Alayrac et al. (2022); Li et al. (2023); Liu et al. (2023a), but most focus on vision-language settings with richer semantic signals and large paired corpora.

2.3 EEG and Large Language Models

Recent work has also begun to connect EEG with language models, including language-supervised EEG representation learning Camaret Ndir et al. (2025), EEG-language foundation modeling Jiang et al. (2025), and LLM-assisted EEG analysis systems Kim et al. (2024); Zhao et al. (2026); Chen et al. (2025); Ma et al. (2026).

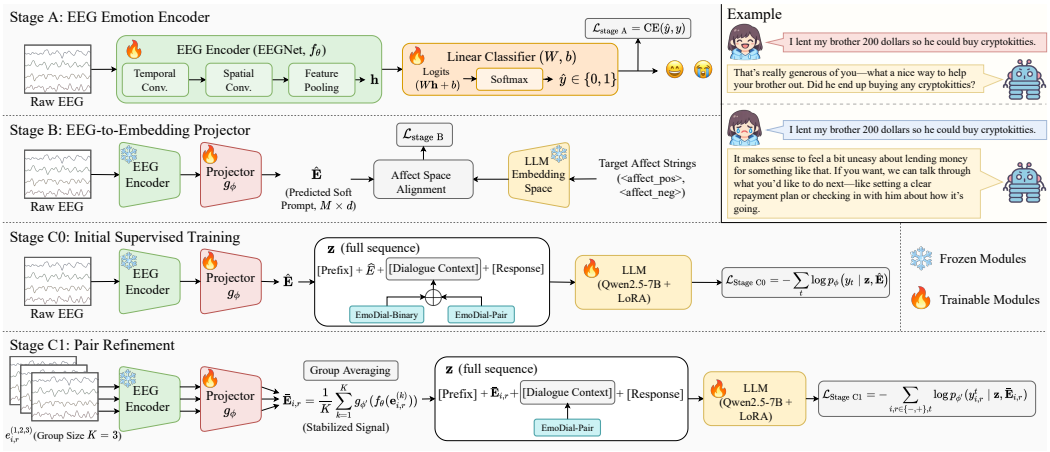


Figure 1: Main pipeline of the proposed EEG-conditioned framework. Stage A learns cross-subject emotion EEG representations, Stage B maps them into the LLM embedding space, and Stage C uses the projected EEG signal for response generation; the upper-right panel shows a counterfactual example of our model.

Some recent work extends EEG+LLM systems to other downstream uses. Thought2Text studies EEG-to-text generation Mishra et al. (2025), while NeuroChat uses EEG-derived feedback to drive a neuroadaptive chatbot for customized learning experiences Baradari et al. (2025). Different from classification or representation learning, our work leverages emotional information in EEG to improve affect-level response adaptation in LLM dialogue generation.

3 Method

Figure 1 summarizes the full pipeline and includes an illustrative counterfactual example in the upper-right panel. The following subsections describe the problem formulation and the pipeline stages.

3.1 Problem Definition

Let $x = \{u_1, \ldots, u_T\}$ denote the dialogue context up to the current turn, where u_t is a user or system utterance, and let e denote an EEG segment reflecting the user's current state. We aim to learn a conditional generator $p(y \mid x, e)$ that produces a response y that is coherent with the dialogue context and better matched to the user's emotional state. At inference time, the model receives no explicit emotion label or instruction, such as "positive" or "negative"; the EEG signal must instead serve as an implicit continuous control signal. The task is therefore not EEG-to-text decoding or standalone EEG emotion classification. Rather, EEG is used to modulate response tone, wording, and supportive strategy.

Because large-scale paired EEG-dialogue data are unavailable, we decompose the problem into three sub-goals: learning cross-subject emotion-related EEG representations, aligning them with the LLM's affect-token embedding space, and using the aligned EEG signal as a soft prompt for response generation. We adopt a cross-subject protocol with non-overlapping train, validation, and test subjects to evaluate generalization to unseen users.

3.2 Stage A: EEG Emotion Encoder

Stage A learns transferable emotion-related EEG features for later alignment with the LLM. We train this encoder on windowed raw EEG segments from the SEED dataset Zheng & Lu (2015) under a cross-subject protocol with strictly non-overlapping train, validation, and test subjects. Let $\mathbf{e} \in \mathbb{R}^{C \times T}$ denote an EEG sample, where C is the number of channels and T is the number of time points in the analysis window. We adopt EEGNet Lawhern et al. (2018) as the Stage-A backbone.

EEGNet combines temporal filtering, spatial filtering across channels, and a compact convolutional feature extractor to produce an emotion-related EEG representation, which is then passed to a linear classifier for binary emotion prediction $y \in \{0, 1\}$, corresponding to NEGATIVE and POSITIVE.

Formally, the encoder produces a latent feature $\mathbf{h} = f_\theta(\mathbf{e})$, which is fed to a linear classifier and optimized with cross-entropy for binary emotion prediction. After training, we discard the classification head and retain the frozen encoder $f_\theta(\cdot)$ as the EEG backbone for Stage B and the later EEG-conditioned generation stages.

3.3 Stage B: EEG-to-Embedding Projector

In Stage B, we freeze the EEG encoder from Stage A as well as the LLM input embedding space, and train only a lightweight projector that maps EEG representations into the embedding sequence of an emotion control token. Let $\mathbf{h}$ denote the frozen EEG feature and let $g(\cdot)$ be the projector. Its output is

$$\hat{\mathbf{E}} = g(\mathbf{h}) \in \mathbb{R}^{M \times d}, \tag{1}$$

where M is the number of soft tokens and d is the LLM embedding dimension. For each emotion label $y \in \{0, 1\}$, we build a frozen target sequence $\mathbf{E}_y \in \mathbb{R}^{M \times d}$ from the input embeddings of the textual control token (<affect_neg> or <affect_pos>). Since the control string may be split into multiple sub-tokens by the tokenizer, we use a binary mask $\mathbf{m}_y \in \{0, 1\}^M$ to mark valid token positions and padding positions.

The Stage B objective aligns the predicted soft prompt with the target affect-token embeddings through token-level matching and prototype-based class regularization. Specifically, it combines masked SmoothL1 and token-wise cosine losses with a prototype classification loss and a margin term in the frozen LLM embedding space. Full term definitions and hyperparameters are given in Appendix C.2.

3.4 Stage C: EEG-Conditioned Response Generation

Stage C trains the dialogue generator in two steps. Stage C0 learns basic EEG-conditioned generation from mixed text supervision, and Stage C1 refines the same model on paired POS/NEG supervision so that the control signal is used more consistently.

3.4.1 Text Supervision Resources

Because there is no large-scale corpus of dialogue responses paired with synchronized EEG, Stage C uses text-side supervision derived from EmpatheticDialogues Rashkin et al. (2019). We construct **EmoDial-Binary** by mapping fine-grained emotions to POS/NEG labels, and **EmoDial-Pair** by creating POS/NEG response pairs for contexts whose text does not strongly reveal polarity. Stage C0 mixes both resources to learn general emotion-aware generation, while Stage C1 uses only EmoDial-Pair to strengthen control sensitivity. Dataset construction details are provided in Appendix A.2.

3.4.2 Stage C0: Initial Supervised Training

We start from the Stage B projector, freeze the EEG encoder, and keep only the projector and the LoRA parameters of the LLM trainable. The projected EEG embeddings replace the original explicit affect-token position and act as a continuous control signal.

Let $P(x)$ denote the fixed tokenized prefix before the EEG soft prompt, $C(x)$ denote the tokenized dialogue context that follows it, and y denote the target response. The full embedded training sequence is constructed as

$$\mathbf{Z} = \Big[\mathrm{Emb}(P(x)), \hat{\mathbf{E}}, \mathrm{Emb}(C(x)), \mathrm{Emb}(y)\Big]. \tag{2}$$

During training, only the response tokens are supervised; the prefix, EEG soft prompt, and dialogue context are masked out in the loss. Stage C0 is optimized with the standard autoregressive language-modeling objective

$$\mathcal{L}_{\text{Stage C0}} = -\sum_{t=1}^{|y|} \log p_\phi\Big(y_t \mid P(x), \hat{\mathbf{E}}, C(x), y_{<t}\Big), \tag{3}$$

where ϕ includes the LoRA parameters of the LLM and the trainable projector parameters.

3.4.3 STAGE C1: PAIR REFINEMENT

Stage C1 continues from the best single-EEG Stage C0 checkpoint. The EEG encoder remains frozen, while the projector and LoRA adapter are further tuned to sharpen the distinction between POS- and NEG-style responses to the same input. Unlike Stage C0, which uses a single label-matched EEG segment, each branch in Stage C1 uses a group of label-matched EEG segments to stabilize the control signal.

Let $\mathcal{D}_{\text{pair}} = \{(x_i, y_i^-, y_i^+)\}_{i=1}^N$ denote the EmoDial-Pair dataset, where x_i is a dialogue context that can plausibly correspond to either negative or positive emotion, and y_i^- and y_i^+ are the corresponding NEG-style and POS-style target responses for that same input. For each pair, we form one negative branch and one positive branch, and pair each branch with a label-matched EEG group $\{\mathbf{e}_{i,r}^{(k)}\}_{k=1}^K$ from the corresponding emotion class. The resulting soft prompts are averaged into a single conditioning sequence,

$$\bar{\mathbf{E}}_{i,r} = \frac{1}{K} \sum_{k=1}^{K} g_{\phi'}(f_\theta(\mathbf{e}_{i,r}^{(k)})), \tag{4}$$

where $r \in \{-, +\}$ and $K = 3$ in our main setting. Stage C1 is still optimized with an autoregressive supervised objective:

$$\mathcal{L}_{\text{Stage C1}} = -\sum_{i=1}^{N} \sum_{r \in \{-,+\}} \sum_{t=1}^{|y_i^r|} \log p_{\phi'}\left(y_{i,t}^r \mid x_i, \bar{\mathbf{E}}_{i,r}, y_{i,<t}^r\right), \tag{5}$$

4 OFFLINE EXPERIMENTAL SETUP

4.1 DATASETS AND SPLITS

For EEG, we use the SEED dataset Zheng & Lu (2015), which contains emotion-elicited recordings from 15 subjects. In the final setting of this paper, we adopt a binary emotion formulation and keep only the POS and NEG conditions. Windowed raw EEG segments with a fixed 4-s analysis window are used to train the Stage A encoder, the Stage B projector, and the label-matched conditioning signals used in Stage C. We use subjects 1–9 for training, 10–12 for validation, and 13–15 for testing.

For text supervision, we use the two Stage C resources introduced in Section 3.4: **EmoDial-Binary** and **EmoDial-Pair**. Their training sizes are 30,379 samples for EmoDial-Binary and 14,960 instances (7,480 POS/NEG pairs) for EmoDial-Pair. Detailed emotion mapping and the pair-construction prompt are provided in Appendices A.1 and A.3, respectively.

For generation evaluation, we further construct a unified 1,598-sample pair-suitable test set from the EmpatheticDialogues test split using the same filtering principle as in pair construction.

4.2 MODELS AND OFFLINE EVALUATION PROTOCOL

All generative variants are built on Qwen2.5-7B-Instruct Team (2024b; 2025a) with LoRA Hu et al. (2022) adapters. Unless otherwise stated, we use the raw-EEG EEGNet setting; in the main configuration, Stage C1 uses $K = 3$.

Our main generation comparison includes four models. **Unconditioned LLM** is the original Qwen2.5-7B-Instruct used without affect control. **Gold-Label Oracle** follows the same training pipeline but receives the ground-truth POS/NEG label at inference time, serving as a reference upper bound for EEG-conditioned generation. **Text-Only Baseline** follows the same training pipeline as the Gold-Label Oracle, but no emotion label is provided at inference time. **EEG-Conditioned LLM (ours)** replaces that control input with the EEG-derived signal. In addition to this four-model comparison, we compare our method with several open-weight LLMs of similar scale under the same evaluation protocol.

Offline evaluation focuses on response generation. We first conduct blind GPT-5.4 evaluation on the 1,598 pair-suitable test samples described in Section 4.1. Candidate responses are anonymized, randomly shuffled, and judged three times per sample. The judge outputs a ranking together with five scores: emotion fit, humanity, helpfulness, naturalness, and overall quality. Among these metrics, ranking and emotion fit are the most important because they most directly reflect the model's ability to use emotion cues to shape the response. The blind-judge prompt is included in Appendix D.3.

We additionally conduct a questionnaire-based human evaluation on the first 500 samples from the same test set to check whether the blind LLM-judge results are consistent with human preference. A total of 24 participants were recruited to complete the questionnaire. For each sample, the anonymized responses from the same four models are evaluated by three human raters. Each rater provides an overall 1–5 score and a ranking.

Additional analyses include ablations on Stage C1 relative to Stage C0, projector-based control versus direct EEG label control, single- versus multi-EEG refinement in Stage C1, and other analyses in Appendix E.

5 Online Real-Time EEG User Study

5.1 Study Goal

In addition to the offline benchmark, we conducted a real-time user study in which each participant's EEG was acquired during interaction and fed to the EEG-Conditioned LLM. Participants were not constrained to dataset prompts and could freely compose their own utterances, allowing us to test the model beyond the offline corpus in a more open interaction setting. This complements the offline benchmark because EEG, dialogue, response, and rating are collected in the same live episode.

5.2 Participants

The study recruited 18 university students. All participants provided written informed consent before the study; demographic and screening details are provided in Appendix B.1.

5.3 Procedure and Evaluation

We used film clips to induce participants' emotional states, following SEED and SEED-IV materials in which candidate videos were manually rated and selected to reliably elicit the target emotions Zheng & Lu (2015); Zheng et al. (2018). Each participant completed 24 trials, including 12 positive and 12 negative clips.

As shown in Figure 2, each trial first presented an emotion-induction video, after which the participant entered a self-composed dialogue utterance. The current EEG and text context were then fed to the models, testing the system beyond fixed dataset prompts. The online comparison includes the EEG-Conditioned LLM, the Unconditioned LLM, and the Text-Only Baseline. For each trial, anonymized responses from all systems were randomly shuffled and presented together. The participant rated each response on a 1–5 emotion-fit scale, and model ranks were derived from these ratings. We report average emotion-fit score and derived average rank. Detailed protocol and implementation settings are provided in Appendices B.1 and B.2.

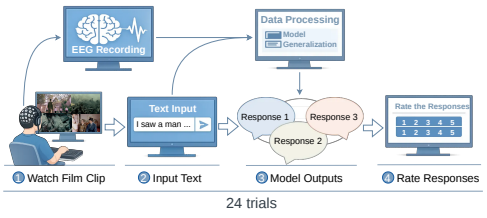


Figure 2: Overview of the online real-time EEG user study.

Table 1: Main blind LLM-judge results on the 1,598-sample test set.

Model	Win Rate (%)↑	Avg Rank↓	Emotion Fit↑	Humanity ↑	Helpfulness ↑	Naturalness ↑	Overall ↑
Unconditioned LLM	11.67	3.17	5.81	6.18	**6.89**	6.44	6.04
Text-Only Baseline	22.44	2.45	6.93	7.19	6.50	**7.93**	7.06
EEG-Conditioned LLM (ours)	**31.93**	**2.23**	**7.49**	**7.60**	6.42	**7.93**	**7.38**
Gold-Label Oracle (reference upper bound)	33.96	2.14	7.69	7.74	6.57	7.89	7.53

Table 2: Comparison with similarly sized peer LLMs.

Model	Params	Win Rate (%)↑	Avg Rank↓	Emotion Fit↑	Humanity ↑	Helpfulness ↑	Naturalness ↑	Overall ↑
Qwen2.5-7B Team (2024b; 2025a)	7B	6.26	3.70	6.34	6.41	**6.89**	6.76	6.38
Llama3-8B AI@Meta (2024)	8B	15.39	3.21	6.67	6.72	6.20	7.41	6.65
Qwen3-8B Team (2025b)	8B	20.21	3.14	6.84	7.03	6.77	7.28	6.87
Gemma-2-9B-it Team (2024a)	9B	19.77	3.14	6.75	6.85	6.24	7.03	6.65
Ministral-3-8B Liu et al. (2026)	8B	5.38	4.95	4.58	4.99	4.98	4.25	4.29
EEG-Conditioned LLM (ours)	7B + adapters	**32.98**	**2.86**	**7.39**	**7.49**	6.02	**7.90**	**7.17**

6 Results and Discussion

6.1 Offline Evaluation Results

6.1.1 Main Comparison with Core Baselines

Table 1 summarizes the main blind LLM-judge results on the unified 1,598-sample test set. Among the practical systems, EEG-Conditioned LLM is the strongest model. It achieves a 31.93% win rate and a 2.23 average rank, outperforming Text-Only Baseline (22.44%, 2.45) and Unconditioned LLM (11.67%, 3.17). It also leads the practical baselines on emotion fit, humanity, naturalness, and overall quality. Additional analyses are provided in Appendix E.3.

Helpfulness is the main exception: longer replies may appear more helpful by offering more detail or suggestions, and Unconditioned LLM produces much longer replies (59.3 words on average versus 30.8 for ours). The Gold-Label Oracle remains the reference upper bound because it uses explicit emotion labels unavailable at inference time; our model narrows this gap and attains a 48.09% pairwise win rate against the oracle. The counterfactual example in Figure 1 shows the same dialogue shifting toward optimistic or supportive wording when only the EEG condition changes.

6.1.2 Comparison with Peer LLMs

We next test whether the gains can be explained by backbone choice alone by comparing against five similarly sized open-weight LLMs without EEG conditioning. All peer models use the same dialogue inputs and blind LLM-judge prompt, while our model uses a 7B backbone with a lightweight EEG projector and LoRA adapters. Table 2 shows that the EEG-conditioned model again ranks first, with the highest win rate (32.98%), best average rank (2.86), and best emotion fit, humanity, naturalness, and overall score. Qwen3-8B and Gemma-2-9B-it are the strongest non-EEG competitors, but remain behind on both ranking metrics and emotion-related quality. Helpfulness is again the exception: our model is rated lower despite leading overall, likely in part because its replies are much shorter than those of the peer models (29.2 vs. 42.4–68.8 words).

Table 3: Questionnaire-based human evaluation on 500 test samples.

Model	Avg Score (×2) ↑	Avg Rank ↓
Unconditioned LLM	6.76	2.53
Text-Only Baseline	6.56	2.71
EEG-Conditioned LLM (ours)	**7.05**	**2.45**
Gold-Label Oracle	7.32	2.31

Table 4: Online user-study results.

Model	Avg Emotion-Fit	Avg Rank
Unconditioned LLM	2.56	2.35
Text-Only Baseline	3.22	1.91
EEG-Conditioned LLM (ours)	**3.44**	**1.74**

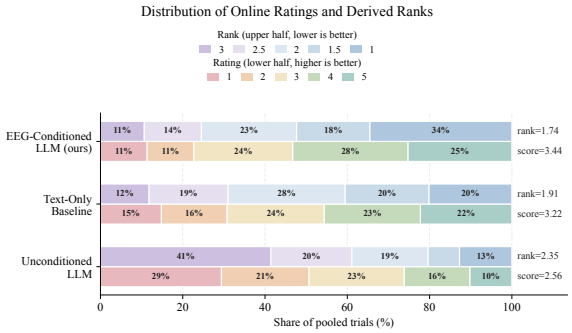


Figure 3: Distribution of online ratings and derived ranks.

6.1.3 Questionnaire Results

Table 3 summarizes the questionnaire-based human evaluation on the first 500 samples from the same 1,598-sample test set. We report average score after rescaling the original 1–5 ratings by 2 to match the 1–10 LLM-judge scale, together with average rank.

The human results are broadly consistent with the blind GPT-5.4 evaluation in Table 1. In both settings, the Gold-Label Oracle remains strongest overall, and EEG-Conditioned LLM is the strongest practical model. The main discrepancy is the Unconditioned LLM: in the questionnaire it slightly exceeds Text-Only Baseline, whereas in blind LLM-judge evaluation the Text-Only Baseline is clearly preferred. A likely reason is that Unconditioned LLM tends to produce longer responses than Text-Only Baseline, and human raters may reward that extra length more than the LLM judge does.

6.2 Online Evaluation Results

Table 4 summarizes the online user-study results aggregated over all trials. EEG-Conditioned LLM again ranks first, with the highest average emotion-fit score (3.44/5) and the best average rank (1.74). Text-Only Baseline ranks second with 3.22 and 1.91, whereas Unconditioned LLM trails at 2.56 and 2.35. Participant-level comparisons are reported in Appendix E.3.

Figure 3 shows the pooled trial-level distributions of derived ranks (upper halves) and emotion-fit ratings (lower halves). EEG-Conditioned LLM receives more high ratings and low ranks, whereas Unconditioned LLM responses are mostly rated 1 or 2.

This ordering agrees with the offline evaluation, while the online study tests a more open setting: EEG is recorded from the current participant and paired with their own utterance, and responses are evaluated by that participant. These user ratings therefore provide complementary evidence that EEG-derived affective control remains useful during live dialogue.

6.3 Ablation Studies

Table 5 evaluates pair refinement, projector-based control, and multi-EEG conditioning.

Stage C1 Relative to Stage C0. Using the same Stage A encoder, Stage B projector configuration, and LLM backbone, Stage C1 wins 69.2% of pairwise judgments against Stage C0, reaches a 1.31 average rank, and improves all five quality scores.

Table 5: Summary of the three main ablations.

Variant	Win Rate (%)↑	Avg Rank↓	Emotion Fit↑	Humanity ↑	Helpfulness ↑	Naturalness ↑	Overall ↑
Stage C1 Relative to Stage C0							
Stage C0 Initialization	30.8	1.69	6.43	6.50	5.38	7.54	6.36
Stage C1 Refinement (ours)	**69.2**	**1.31**	**7.78**	**7.78**	**6.75**	**8.00**	**7.72**
Projector-Based Control versus Direct EEG Label Control							
Direct EEG Label Control	43.6	1.56	7.13	7.43	6.47	7.78	7.18
Projector-Based EEG Control (ours)	**56.4**	**1.44**	**7.57**	**7.54**	**6.63**	**7.91**	**7.49**
Single- versus Multi-EEG Refinement in Stage C1							
Single-EEG Stage C1 ($K = 1$)	47.4	1.53	7.29	7.48	6.47	**7.91**	7.32
Multi-EEG Stage C1 ($K = 3$, ours)	**52.6**	**1.47**	**7.46**	**7.55**	**6.58**	7.87	**7.42**

Projector-Based Control versus Direct EEG Label Control. We compare the projector with explicit POS/NEG control from EEG classification, sharing the Stage A encoder, Stage B initialization, Stage C supervision, and LLM backbone. Projector-based control wins 56.4% of pairwise judgments, achieves a 1.44 average rank, and improves all five quality scores. One possible explanation is that projection retains a continuous control signal, whereas hard labels compress EEG information and pass classification errors directly to the generator.

Single- versus Multi-EEG Refinement in Stage C1. Compared with single-EEG conditioning ($K = 1$), the main multi-EEG setting ($K = 3$) wins 52.6% of pairwise judgments and achieves a 1.47 average rank, with modest gains in emotion fit, humanity, helpfulness, and overall quality. Additional results for $K \in \{1, 2, 3, 5\}$ are in Appendix E.6.

6.4 Discussion

We also examine whether the model changes its outputs when only the EEG condition changes. With the dialogue context fixed and only the EEG condition swapped, the final EEG-conditioned model produces exactly identical responses for only 7.2% of pairs, and only 8.4% are near-duplicates with very high lexical overlap. This counterfactual check suggests that the model uses EEG as a control signal rather than producing fixed responses regardless of the input EEG.

These findings are still limited by label-matched but non-synchronized EEG-text training data, the binary POS/NEG setting, and reliance on LLM judging for the main offline evaluation. In addition, part of the Stage C supervision is derived from GPT-generated POS/NEG response pairs. Therefore, the present evidence is better interpreted as supporting affect-level EEG control rather than fine-grained neural-language alignment. The questionnaire and online study show the same overall ordering under human evaluation, but they do not fully remove possible judge-side stylistic bias. Future work should pursue synchronized EEG-dialogue corpora with turn-level neural and language alignment, extend the control space beyond binary polarity to richer affective states, and explore broader physiological or user-level adaptation.

7 Conclusion

We introduced a staged framework that uses emotion EEG as an implicit control signal for affect-level response adaptation in LLM-based dialogue. It combines cross-subject EEG encoding, projection into the LLM embedding space, and soft-prompt conditioning without explicit emotion labels at inference time. Blind LLM-judge evaluation, a 500-sample human questionnaire, and a preliminary real-time user study consistently favored EEG-conditioned generation over text-only baselines. Ablations supported the contributions of Stage C1 refinement, projector-based control, and multi-EEG conditioning. Together, these findings support EEG as an implicit signal for adapting dialogue responses to users' emotional states.

AI USE STATEMENT

Generative AI tools assisted with language polishing, the search and discovery of relevant literature, and research implementation, particularly code development and debugging. They were also used to prepare initial versions of the synthetic text datasets in this study. References suggested by AI were verified against the original publications. The authors reviewed and revised AI-assisted prose, code, and synthetic text before incorporating them into the work, and take full responsibility for the manuscript, code, data, and scientific claims.

ETHICS STATEMENT

This study used the publicly available SEED emotion EEG dataset Zheng & Lu (2015) and included an online EEG experiment with recruited participants. All participants in our experiment signed written informed consent forms before taking part and received monetary compensation for their participation.

REPRODUCIBILITY STATEMENT

We will publicly release all code upon acceptance of the paper. The model architecture and experimental protocols are described in the main paper, with training configurations and implementation details in Appendix C and data preprocessing, software environment, and reproducibility details in Appendix G.

REFERENCES

AI@Meta. Llama 3 model card. `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion recognition based on EEG using LSTM recurrent neural network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017.

Dünya Baradari, Nataliya Kosmyna, Oscar Petrov, Rebecah Kaplun, and Pattie Maes. NeuroChat: A neuroadaptive AI chatbot for customizing learning experiences. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–21, 2025.

Tidiane Camaret Ndir, Robin T Schirrmeister, and Tonio Ball. EEG-CLIP: Learning EEG representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025.

Hongyu Chen, Weiming Zeng, Chengcheng Chen, Luhui Cai, Fei Wang, Yuhu Shi, Lei Wang, Wei Zhang, Yueyang Li, Hongjie Yan, et al. EEG emotion copilot: Optimizing lightweight LLMs for emotional EEG interpretation with assisted medical record generation. *Neural Networks*, pp. 107848, 2025.

Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin JunLiang Lim, and Cuntai Guan. EmT: A novel transformer for generalized cross-subject EEG emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *The Tenth International Conference on Learning Representations*, 1(2):3, 2022.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. SEED-VII: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2024.

Weibang Jiang, Yansen Wang, Bao-liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jonathan W Kim, Ahmed Alaa, and Danilo Bernardo. EEG-GPT: exploring capabilities of large language models for EEG classification and interpretation. *arXiv preprint arXiv:2401.18006*, 2024.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

He Li, Yi-Ming Jin, Wei-Long Zheng, and Bao-Liang Lu. Cross-subject emotion recognition using deep adaptation networks. In *International Conference on Neural Information Processing*, pp. 403–413. Springer, 2018.

Jingcong Li, Shuqi Li, Jiahui Pan, and Fei Wang. Cross-subject EEG emotion recognition with self-organized graph neural network. *Frontiers in Neuroscience*, 15:611653, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. EmpDG: Multiresolution interactive empathetic dialogue generation, 2019.

Rui Li, Yiting Wang, and Bao-Liang Lu. Measuring decision confidence levels from EEG using a spectral-spatial-temporal adaptive graph convolutional neural network. In *International Conference on Neural Information Processing*, pp. 395–406. Springer, 2022a.

Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6–14, 2022b.

Alexander H. Liu et al. Ministral 3, 2026. URL `https://arxiv.org/abs/2601.08584`.

Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. Speak from heart: an emotion-guided LLM-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 533–542, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023b.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

Bo-Qun Ma, He Li, Wei-Long Zheng, and Bao-Liang Lu. Reducing the subject variability of EEG signals with adversarial domain generalization. In *International Conference on Neural Information Processing*, pp. 30–42. Springer, 2019.

Fei Ma, Han Lin, Yifan Xie, Hongwei Ren, Xiaoyu Shen, Wenbo Ding, and Qi Tian. e^2-LLM: Bridging neural signals and interpretable affective analysis. *arXiv preprint arXiv:2601.07877*, 2026.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking emotions for empathetic response generation, 2020.

Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2Text: text generation from EEG signal using large language models (LLMs). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3747–3759, 2025.

Yushan Qian, Weinan Zhang, and Ting Liu. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6516–6528, 2023.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, 2019.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1): 42–55, 2011.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2018.

Gemma Team. Gemma. `https://www.kaggle.com/m/3301`, 2024a.

Qwen Team. Qwen2 technical report, 2024b. URL `https://arxiv.org/abs/2407.10671`.

Qwen Team. Qwen2.5 technical report, 2025a. URL `https://arxiv.org/abs/2412.15115`.

Qwen Team. Qwen3 technical report, 2025b. URL `https://arxiv.org/abs/2505.09388`.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4634–4645, 2022.

Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2202–2218, 2025.

Sha Zhao, Mingyi Peng, Haiteng Jiang, Tao Li, and Shijian Li. EEG agent: A unified framework for automated EEG analysis using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18063–18071, 2026.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

Wei-Long Zheng and Bao-Liang Lu. Personalizing EEG-based affective models with transfer learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2732–2738, 2016.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2018.

Table 6: Emotion mapping and target response styles.

Original emotion categories / cases	Binary label	Target response style
excited, proud, grateful, confident, hopeful, impressed, joyful, content, caring, trusting, faithful	POS	acknowledging, encouraging, celebrating, sharing positive affect
angry, annoyed, sad, afraid, lonely, terrified, guilty, anxious, disgusted, furious, disappointed, jealous, devastated, embarrassed, ashamed, apprehensive	NEG	validating feelings, comforting, reassuring, providing support
nostalgic, sentimental, surprised, anticipating, prepared	removed	not used for binary training because the polarity is ambiguous for the current binary setting

Peixiang Zhong, Di Wang, and Chunyan Miao. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A Text-Side Supervision Data

A.1 Emotion Mapping and Emotional Response Styles

Table 6 specifies how the original EmpatheticDialogues emotion categories were converted into the binary POS/NEG setting. The same mapping was used to derive text-supervision labels and to define the target response style associated with each affective condition.

This mapping makes the control objective explicit. Positive states are paired with acknowledging and encouraging responses, while negative states are paired with validation, comfort, and reassurance. Ambiguous categories such as *nostalgic* or *surprised* were removed because their polarity is less stable in the present binary formulation.

A.2 Construction Statistics for the Text-Side Supervision

Table 7 summarizes the text-side datasets used in the main paper and their sizes: **EmoDial-Binary** contains 30,379 training instances, **EmoDial-Pair** contains 14,960 instances constructed as 7,480 POS/NEG pairs, and the unified **Pair-Suitable Test Set** contains 1,598 instances constructed as 799 pairs.

The source data are derived from EmpatheticDialogues, which provides dialogue contexts, responses, and 32 fine-grained emotion labels. We map clearly negative emotions to NEG and clearly positive emotions to POS, while removing ambiguous categories. For paired supervision, we focus on contexts whose text does not strongly reveal emotional polarity, use GPT-5.3 to identify suitable cases, and generate one POS-style and one NEG-style response for each selected context. Stage C0 mixes EmoDial-Binary and EmoDial-Pair at a source-aware ratio of $0.6 : 0.4$, and Stage C1 uses only EmoDial-Pair.

A.3 Prompt for Pair Data Construction

System prompt.

Table 7: Text-side dataset sizes used in the main paper.

Dataset	Size
EmoDial-Binary	30,379
EmoDial-Pair	14,960
Pair-Suitable Test Set	1,598

```
You are an expert curator for emotional dialogue training data. Given a
    dialogue input, the user's current emotion label, and a reference
    response, decide whether the sample is suitable for contrastive POS/
    NEG rewriting, then generate high-quality paired responses. Always
    return strict JSON only.
```

User prompt template. Instruction and suitability block.

```
Task:
1) Decide if this sample is suitable for differential emotional rewriting
    .
2) If suitable, generate two assistant replies with consistent factual
    content but different emotional style:
   - POS: shared positive affect, encouraging tone.
   - NEG: validation, comfort, emotionally supportive tone.

Suitability rule (strict):
- Reject ONLY when the user explicitly self-reports strong emotion (very
    obvious), such as "I am very happy" or "I feel very sad".
- If user emotion is not explicitly very obvious, set suitable=true and
    generate both POS/NEG replies.
- Also set suitable=false if style change would distort core meaning.
```

Generation and scoring block.

```
Constraints for generated replies:
- Keep one-turn assistant response only.
- Keep key facts/intents consistent with the reference.
- No role tags like 'Assistant:'.
- Keep each reply to 1-3 sentences.

Scoring rule for input emotion tendency (0-100):
- 0: no clear tendency.
- 30-60: some weak tendency clues.
- 80+: very obvious, explicit self-report of strong emotion.
```

Output schema block.

```
Return JSON schema exactly:
{
  "input_emotion_tendency": "very_obvious_positive|very_obvious_negative|
    not_obvious",
  "input_emotion_tendency_score": 0-100 integer,
  "suitable": true/false,
  "reason": "...",

  "confidence": 0-100 integer,
  "pos_reply": "...",
  "neg_reply": "..."
}
```

Injected fields block.

```
Given source user emotion label from dataset:
<SOURCE_EMOTION>
```

```
Local pre-check tendency (for reference only):
<LOCAL_TENDENCY>


Extracted user-signal text for tendency judgement:
<<<
<SIGNAL_TEXT>
>>>

Dialogue input:
<<<
<INPUT_TEXT>
>>>


Reference assistant reply:
<<<
<REF_OUTPUT>
>>>
```

B Online User Study Details

B.1 Protocol

The online study involved 18 university students (9 female and 9 male), aged 19 to 32 years (mean 23.94, standard deviation 2.97). Participants reported no history of neurological disorders, psychiatric conditions, or substance abuse, and all provided written informed consent before the study. Each participant received fixed monetary compensation.

We used film clips to induce participants' emotional states, following SEED and SEED-IV materials in which candidate videos were manually rated and selected to reliably elicit the target emotions Zheng & Lu (2015); Zheng et al. (2018). Each participant completed 24 trials, including 12 positive and 12 negative clips. In each trial, the system presented three anonymized model responses in random order, and the participant rated each response on a 5-point scale. The per-trial model ranks were derived from these within-trial ratings, with tied scores assigned average ranks.

The online experiment used a custom PsychoPy interface connected to a real-time EEG inference server. The online acquisition used the same 62-channel wet-electrode EEG setup as the offline pipeline. Each trial began with an emotion-induction video segment, after which the participant freely composed a dialogue utterance. The three-machine setup separated EEG acquisition/control, the participant interface, and model inference, so acquisition remained stable while the inference workstation loaded the LLM and returned randomized candidate responses for rating.

B.2 Online Inference Pipeline and Runtime

Following the same preprocessing logic as in offline training, the online server extracted a 4-second EEG window, applied a 75 Hz low-pass filter, and resampled the signal to 200 Hz. The resulting EEG tensor and the participant's text input were then fed to the EEG-conditioned model, while the same text input was also provided to the text-only comparison systems under identical decoding settings. No participant-specific calibration or fine-tuning was performed during the online study. The same frozen EEG encoder and projector were used for all participants, matching the cross-subject generalization setting used in the offline experiments.

From a systems perspective, the online pipeline incurred a one-time startup cost at the beginning of a session to load the LLM, LoRA adapter, and EEG modules onto the GPU. After this initialization, the per-trial latency was dominated by LLM decoding and lightweight transmission between the acquisition interface and the inference server. Relative to text generation, EEG preprocessing and the projector forward pass added little additional delay, so the interaction remained responsive from the participant's perspective.

Table 8: Preset Sweep Space for Stage A and Stage B

Component	Hyperparameter	Searched values / presets	Final adopted value
Stage A (raw EEGNet)	architecture and optimization	11 preset combinations over F1 $\in \{4, 8, 16\}$, D $\in \{2, 3\}$, F2 $\in \{8, 16, 32, 48\}$, kernel length $\in \{64, 128\}$, dropout $\in \{0.35, 0.40, 0.45, 0.50, 0.60\}$, learning rate $\in \{5 \times 10^{-4}, 7 \times 10^{-4}, 10^{-3}\}$, and weight decay $\in \{0, 10^{-4}, 3 \times 10^{-4}\}$; all presets used batch size 64, patience 40, and at most 300 epochs	`long_kernel`: F1= 8, D= 2, F2= 16, kernel length = 128, dropout = 0.5, learning rate $= 7 \times 10^{-4}$, weight decay $= 10^{-4}$, batch size = 64
Stage B (projector)	projector width and regularization	6 preset combinations over projector hidden dimension $\in \{384, 512, 768, 1024, 1536\}$, projector dropout $\in \{0.10, 0.15, 0.20, 0.25, 0.30\}$, learning rate $\in \{3 \times 10^{-4}, 5 \times 10^{-4}, 7 \times 10^{-4}, 10^{-3}\}$, and weight decay $\in \{10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}\}$; all presets used batch size 256, feature batch size 256, patience 8, at most 30 epochs, and loss mode `v2_lite`	`wide_lowlr`: hidden dim = 1536, projector dropout = 0.15, learning rate $= 3 \times 10^{-4}$, weight decay $= 2 \times 10^{-2}$, batch size = 256

Table 9: Held-out test performance of the final Stage A encoder.

Metric	Value
Accuracy	0.8302
Balanced accuracy	0.8291
Macro-F1	0.8294
Weighted-F1	0.8297
NEG recall	0.7790
POS recall	0.8793

C Modeling, Training, and Inference Details

C.1 Stage A: Raw EEG Encoder Search Space and Final Performance

C.1.1 Hyperparameter Sweep for the Raw EEG Encoder and Projector

Table 8 summarizes the preset sweep space used for the raw-EEG branch in Stages A and B. Rather than tuning each hyperparameter independently, we grouped them into a small number of presets so that the cross-subject model-selection process remained tractable. For Stage A, the selected `long_kernel` preset uses a relatively long temporal kernel and moderate channel width, which gave the strongest cross-subject affect representation in our sweep. For Stage B, the selected `wide_lowlr` preset uses a wider projector with a lower learning rate, which gave more stable mapping from EEG features to affect-token embeddings.

C.1.2 Final Stage A Encoder Result

Table 9 reports the held-out cross-subject classification result of the final Stage A raw EEGNet encoder adopted in the main pipeline. This is the `long_kernel` preset selected from the search space in Appendix C.1.1. These numbers show that the final encoder learns a stable binary affect representation before the later projector and generation stages, with both classes recognized well and somewhat higher recall on the POS side.

Table 10: Stage C training and decoding configuration.

Component	Setting
Backbone	Qwen2.5-7B-Instruct
Stage C0	1 epoch; batch size 2; gradient accumulation 16; learning rate 2×10^{-5}; projector learning rate 10^{-4}; validation ratio 0.03; LoRA rank 16, alpha 32, dropout 0.05
Stage C1	1 epoch; batch size 2; gradient accumulation 16; learning rate 2×10^{-6}; projector learning rate 10^{-5}; multi-EEG refinement with $K = 3$ in the main setting and mean-projected group reduction
Decoding	Greedy decoding (`do_sample=false`); maximum 128 new tokens; repetition penalty 1.0

C.2 Stage B: EEG-to-Affect Soft-Prompt Projection Objective

This section gives the full term definitions for the Stage B objective in the main paper. For a sample with affect label $y \in \{0, 1\}$, the projector predicts a soft-prompt embedding sequence $\hat{\mathbf{E}} \in \mathbb{R}^{M\times d}$, and $\mathbf{E}_y \in \mathbb{R}^{M\times d}$ is the frozen target affect-token sequence with binary mask $\mathbf{m}_y \in \{0, 1\}^M$.

We first apply a masked token-level SmoothL1 loss:

$$\mathcal{L}_{\text{token}} = \frac{1}{\|\mathbf{m}_y\|_1 d} \sum_{t=1}^{M} m_{y,t}\, \text{SmoothL1}\Big(\hat{\mathbf{E}}_t, \mathbf{E}_{y,t}\Big). \tag{6}$$

We also use a token-wise cosine alignment term:

$$\mathcal{L}_{\text{tokcos}} = \frac{1}{\|\mathbf{m}_y\|_1} \sum_{t=1}^{M} m_{y,t} \left(1 - \cos\Big(\hat{\mathbf{E}}_t, \mathbf{E}_{y,t}\Big)\right). \tag{7}$$

For class-level regularization, we first pool the predicted sequence into

$$\hat{\mathbf{p}} = \frac{\sum_{t=1}^{M} m_{y,t}\hat{\mathbf{E}}_t}{\sum_{t=1}^{M} m_{y,t}}, \tag{8}$$

and define the prototype of class k as

$$\mathbf{c}_k = \frac{\sum_{t=1}^{M} m_{k,t}\mathbf{E}_{k,t}}{\sum_{t=1}^{M} m_{k,t}}, \qquad k \in \{0, 1\}. \tag{9}$$

Prototype logits are computed by temperature-scaled cosine similarity:

$$s_k = \frac{\langle \hat{\mathbf{p}}/\|\hat{\mathbf{p}}\|_2,\ \mathbf{c}_k/\|\mathbf{c}_k\|_2 \rangle}{\tau}. \tag{10}$$

This gives a prototype classification loss with label smoothing,

$$\mathcal{L}_{\text{protoCE}} = \text{CE}_\epsilon(\mathbf{s}, y), \tag{11}$$

and a margin term,

$$\mathcal{L}_{\text{margin}} = \max\left(0, \gamma - s_y + \max_{k\neq y} s_k\right). \tag{12}$$

The full Stage B objective is

$$\mathcal{L}_{\text{Stage B}} = \lambda_1\mathcal{L}_{\text{token}} + \lambda_2\mathcal{L}_{\text{tokcos}} + \lambda_3\mathcal{L}_{\text{protoCE}} + \lambda_4\mathcal{L}_{\text{margin}}. \tag{13}$$

In the main experiments, we use $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$, $\lambda_4 = 0.1$, $\tau = 0.1$, $\epsilon = 0.05$, $\gamma = 0.1$, and SmoothL1 parameter $\beta = 1.0$.

C.3 Stage C: LLM Adaptation and Decoding Configuration

Table 10 summarizes the fixed Stage C training and decoding settings used in the main experiments.

D Evaluation Protocol

D.1 Blind LLM-Judge Configuration

For the main blind evaluation, we used GPT-5.4 as the judge model. For each test sample, the candidate responses were anonymized and evaluated under three independent blind shuffles to reduce ordering effects. The judge was asked to rank all candidates from best to worst and to assign per-response scores along five quality dimensions: emotion fit, humanity, helpfulness, naturalness, and overall quality. Ranking and emotion fit are the primary metrics because they most directly reflect whether the model uses emotion cues to shape its response.

D.2 Metric Definitions and Aggregation

For the blind LLM-judge evaluation, each model receives one rank and five numeric scores for each judged sample under each blind shuffle. We aggregate the three shuffles by averaging the resulting per-model ranks, first-place indicators, and scores. **Win Rate** is the mean first-place indicator, i.e., the percentage of judged comparisons in which a model is ranked first. **Avg Rank** is the mean ordinal rank, where 1 is best and lower values indicate better preference. **Emotion Fit** measures whether the response matches the target affective state and the speaker's emotional needs. **Humanity** measures warmth, empathy, and human-like interaction quality. **Helpfulness** measures whether the response provides useful support without being inappropriate or overly forceful. **Naturalness** measures fluency and conversational plausibility. **Overall** is the judge's holistic quality score for the response in context, rather than a computed average of the other dimensions. All five score metrics use a 1–10 scale.

For the questionnaire-based human evaluation, each rated response receives an overall 1–5 score and a rank among the four anonymized model responses for the same item. The reported **Avg Score** is the mean human score across items and raters, rescaled by $\times 2$ only to place it on the same numeric range as the 1–10 LLM-judge scores. The reported **Avg Rank** is the mean human-assigned rank, where lower values indicate stronger human preference.

D.3 Prompt for LLM-Based Evaluation

This subsection reports the prompt template used by the blind LLM judge. Dynamic fields are shown as placeholders.

D.3.1 System Prompt

This prompt defines the evaluator role and requires strict JSON output.

```
You are an impartial evaluator for empathetic dialogue quality. You will
    receive one dialogue context and multiple candidate assistant
    responses. Do blind evaluation only based on text quality and
    emotional appropriateness. Always output strict JSON only.
```

D.3.2 User Prompt Template

For readability, we present the template in four contiguous blocks below.

Task block.

```
Evaluation task:
Rank <N_OPTIONS> candidate responses from best to worst for this speaker.
Speaker emotion label is provided and must be respected.
```

Scoring block.

```
Scoring dimensions (1-10):
- emotion_fit: how well response fits speaker emotion and emotional needs
    .
- humanity: warmth, empathy, and human-like interaction quality.
- helpfulness: useful support without being pushy.
```

Table 11: Output statistics for the four main models.

Model	Avg Chars	Avg Words	Median Words	Avg Jaccard	Exact Match Rate
Unconditioned LLM	348.16	59.30	53	1.0000	1.0000
Text-Only Baseline	155.95	29.63	30	1.0000	1.0000
EEG-Conditioned LLM (ours)	158.84	30.80	30	0.3265	0.0663
Gold-Label Oracle	169.24	32.52	32	0.2180	0.0000

```
- naturalness: fluency and conversational naturalness.
- overall: holistic quality for this context.
```

Rules and output schema block.

```
Rules:
- No ties in ranking.
- ranking must include all labels exactly once.
- winner must be one label in ranking and equal to the first ranked label
    .
- Valid labels are only: <VALID_LABELS>
- Keep reason concise (<= 80 words).

Return JSON schema exactly:
{
  "ranking": <LABELS_JSON>,
  "winner": "<FIRST_LABEL>",
  "confidence": 0-100,
  "reason": "...",
  "scores": {
    "<LABEL>": {"emotion_fit": 1-10, "humanity": 1-10, "helpfulness":
    1-10, "naturalness": 1-10, "overall": 1-10}
  }
}
```

Injected fields block.

```
Speaker emotion label: <EMOTION_LABEL>

Dialogue context:
<<<
<DIALOGUE>
>>>

Candidate responses (blind):
<<<
<CANDIDATES_BLOCK>
>>>
```

D.4 Questionnaire and Human Rating Setup

In the human evaluation, we used a questionnaire-based protocol on 500 sampled test items. Each item received three human ratings, and the full study involved 24 participants in total. For each item, the anonymized responses from the same four models were evaluated by three raters. Each rater provided an overall 1–5 score and a ranking. These ratings were used as the human-evaluation results in the main experiments.

Table 12: Output statistics for the peer-model comparison.

Model	Avg Chars	Avg Words	Median Words	Avg Jaccard	Exact Match Rate
Qwen2.5-7B	348.16	58.98	53	1.0000	1.0000
Llama3-8B	248.64	44.78	39	1.0000	1.0000
Qwen3-8B	316.38	56.03	51	1.0000	1.0000
Gemma-2-9B-it	241.22	42.37	36	1.0000	1.0000
Ministral-3-8B-Instruct-2512	420.23	68.81	75	1.0000	1.0000
EEG-Conditioned LLM (ours)	158.84	29.17	29	0.2848	0.0663

E Additional Evaluation Analyses

E.1 Basic Statistics of Model Outputs

Table 11 summarizes output statistics for the four main models, while Table 12 reports the corresponding statistics for the peer-model comparison. We include response length, lexical overlap, and exact-match rate to help interpret helpfulness and counterfactual behavior across models.

Here, **Avg Chars** and **Avg Words** denote the mean response length in characters and words, respectively, while **Median Words** denotes the median response length in words. **Avg Jaccard** is the mean word-level Jaccard overlap between the two counterfactual responses generated for the same dialogue when only the conditioning signal is changed, and **Exact Match Rate** is the fraction of such pairs that are identical verbatim.

For the counterfactual check discussed in the main paper, we keep the dialogue context fixed and swap only the EEG condition. The final EEG-conditioned model produces exactly identical responses for 7.2% of pairs, while 8.4% of pairs show very high lexical overlap (word-level Jaccard similarity above 0.9). The mean Jaccard overlap is 0.305 and the median is 0.205.

E.2 Additional Notes on Offline Evaluation Results

The main blind LLM-judge table compares practical systems with a reference upper bound. The Gold-Label Oracle receives explicit POS/NEG labels at inference time and is therefore not a deployable baseline in the intended setting; it indicates the approximate performance of the same training pipeline when the emotion signal is perfect. The EEG-conditioned model does not surpass this oracle, but narrows the gap while requiring only EEG and dialogue context at inference time.

Emotion fit and rank are the most direct indicators of EEG-based affect control. Helpfulness behaves differently because it is partly confounded with response length: longer replies may appear more helpful because they offer more detail or suggestions. In the main comparison, the Unconditioned LLM produces much longer responses than EEG-Conditioned LLM and Text-Only Baseline, which can make its replies appear more helpful even when they are less emotion-matched.

The peer-LLM comparison is intended to rule out the simpler explanation that the gain comes only from the Qwen2.5-7B backbone. The peer models are evaluated under the same dialogue inputs and blind-judge prompt, without EEG conditioning or task-specific adaptation. The EEG-conditioned model uses a 7B-scale backbone with lightweight projector and LoRA adaptation, and still leads on win rate, average rank, emotion fit, humanity, naturalness, and overall quality. This supports the interpretation that the EEG-derived control signal, rather than only backbone choice, contributes to the observed gains.

Additional no-emotion variants in Appendix E.4 provide another check: text-trained systems without an explicit or EEG-derived emotion signal remain competitive on generic quality dimensions, but do not match the EEG-conditioned model on ranking and emotion-related metrics.

E.3 Statistical Testing Details

All inferential tests follow the natural paired unit of the corresponding analysis. For the offline comparison, the unit is the dialogue sample after aggregating the three blind judge runs. For the online

Table 13: Additional statistical details for the main comparisons.

Comparison	Unit	Procedure	Summary
Offline main result: EEG-Conditioned LLM vs. Text-Only Baseline	Dialogue sample ($n = 1{,}598$)	Exact paired sign test on aggregated ranks; paired bootstrap 95% confidence intervals	Sign test $p = 1.69 \times 10^{-6}$; rank advantage $[0.14, 0.29]$; overall-score gain $[0.22, 0.43]$; emotion-fit gain $[0.44, 0.67]$
Online study: EEG-Conditioned LLM vs. Unconditioned LLM	Participant ($n = 18$)	Two-sided Wilcoxon signed-rank test; subject-level bootstrap 95% confidence intervals	Emotion-fit gain 0.88 with 95% CI $[0.50, 1.27]$, $p = 5.0 \times 10^{-4}$; rank advantage 0.61 with 95% CI $[0.34, 0.85]$, $p = 8.4 \times 10^{-4}$
Online study: EEG-Conditioned LLM vs. Text-Only Baseline	Participant ($n = 18$)	Subject-level bootstrap 95% confidence interval	Emotion-fit gain 0.22 with 95% CI $[0.02, 0.43]$

Table 14: Blind LLM-judge results for additional text-trained no-emotion variants.

Model	Win Rate (%)↑	Avg Rank↓	Emotion Fit↑	Humanity ↑	Helpfulness ↑	Naturalness ↑	Overall ↑
Qwen2.5-7B Text-Only	15.37	3.46	6.84	7.20	**6.61**	7.63	6.90
Qwen3-8B Text-Only	16.30	3.36	6.92	7.15	6.44	7.78	6.95
Llama3-8B Text-Only	13.09	3.62	6.69	6.78	5.83	**7.80**	6.63
Gemma-2-9B-it Text-Only	12.06	3.79	6.52	6.57	5.73	7.76	6.47
Ministral-3-8B Text-Only	13.43	3.52	6.82	7.09	6.46	7.68	6.86
EEG-Conditioned LLM (ours)	**29.75**	**3.21**	**7.26**	**7.55**	6.45	7.61	**7.13**

study, the unit is the participant after averaging across that participant's 24 trials. Questionnaire results are reported descriptively from 500 samples with three human ratings per sample.

Table 13 summarizes the statistical comparisons for the offline and online results. In the offline comparison with Text-Only Baseline, the paired sign test and positive confidence intervals support improvements in rank, overall quality, and emotion fit. In the online study, the score and rank gains over Unconditioned LLM are significant under the two-sided Wilcoxon signed-rank tests. The emotion-fit gain over Text-Only Baseline also has a positive confidence interval.

E.4 Additional Comparison with Text-Trained No-Emotion Variants

Table 14 reports an additional comparison under the same blind LLM-judge protocol. Several open-weight backbones are adapted with the same label-supervised text training recipe and then evaluated without emotion input at inference. The Qwen2.5-7B text-only variant corresponds to the Text-Only Baseline in the main experiments, while the other variants provide comparable no-emotion references under different backbones.

The EEG-Conditioned LLM remains strongest within this family of text-trained no-emotion variants. It achieves the highest win rate, the best average rank, and the strongest emotion-fit, humanity, and overall scores. The text-only variants remain competitive on helpfulness and naturalness, but none matches the EEG-Conditioned model on affect-related quality.

E.5 Pairwise Preference Matrix

The main tables report aggregate ranks and scores. Figure 4 unpacks the main four-model comparison into head-to-head preference probabilities. Each off-diagonal cell indicates how often the model on the y-axis is preferred over the model on the x-axis, providing a direct view of pairwise dominance patterns beyond the aggregate summaries in the main paper.

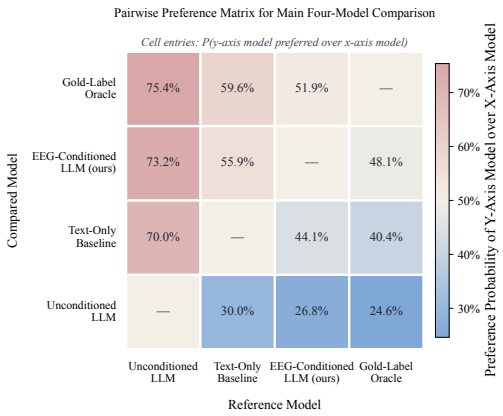


Figure 4: Pairwise preference matrix for the main four-model comparison.

Table 15: Blind LLM-judge results for different Stage C1 EEG aggregation sizes K.

Variant	Avg Rank↓	Emotion Fit↑	Humanity↑	Helpfulness↑	Naturalness↑	Overall↑
$K = 1$	1.97	7.36	7.55	6.59	7.90	7.36
$K = 2$	1.87	7.52	7.63	6.68	7.95	7.49
$K = 3$	1.92	7.48	7.60	6.63	7.88	7.44
$K = 5$	1.92	7.50	7.59	6.66	7.86	7.43

E.6 Sensitivity to the Multi-EEG Group Size

We analyze the sensitivity of the Stage C1 refinement process to the number of label-matched EEG instances aggregated before soft-prompt injection. We vary the aggregation size over $K \in \{1, 2, 3, 5\}$. Here, $K = 1$ acts as the single-EEG reference, and $K = 3$ is the default setting. All settings share the same raw-EEG EEGNet branch, the same cross-subject subject split, and the same evaluation protocol.

The blind LLM-judge results in Table 15 show that the multi-EEG settings $K = 2, 3, 5$ are broadly comparable. All three are slightly better than the single-EEG reference $K = 1$. This supports the interpretation in the main text: the main gain comes from moving from single-EEG to multi-EEG refinement, while the exact choice among $K = 2, 3, 5$ has only a limited effect on judged response quality.

E.7 Representative Case Study

Table 16 provides one representative qualitative case. The dialogue context is fixed across systems, so the comparison isolates how different conditioning strategies change the tone and emotional direction of the response.

F Additional Baseline Experiments

This appendix reports two additional baseline analyses: specialized text-only empathetic dialogue systems and an EEG foundation-encoder hard-label pipeline. These analyses complement the matched text-only and gold-label controls in the main paper.

F.1 Specialized Text-Only Empathetic Dialogue Baselines

To compare against specialized text-only emotion-aware dialogue models, we evaluated EmpDG Li et al. (2019) and MIME Majumder et al. (2020) on the matched subset where both systems produced outputs. MIME has slightly fewer usable examples because no public checkpoint was available; we

Table 16: A representative case.

Field	Content	
Dialogue	User: I lent my brother 200 dollars so he could buy cryptokitties.	
Emotion label	NEG	POS
Unconditioned LLM	That sounds like an interesting loan! Cryptokitties can be quite addictive and fun, but they can also be quite expensive. Did your brother enjoy his experience with them? Have you heard back from him about the payment?	
Text-Only Baseline	That's a lot to lend for something like Cryptokitties—did you end up feeling okay about it? If you're worried, you could gently check in with him about when he'll be able to pay you back.	
EEG-Conditioned LLM (ours)	It makes sense to feel a bit uneasy about lending money for something like that. If you want, we can talk through what you'd like to do next—like setting a clear repayment plan or checking in with him about how it's going.	That's really generous of you—what a nice way to help your brother out. Did he end up buying any cryptokitties?
Gold-Label Oracle	That sounds like a lot to lend, and it makes sense if you're feeling unsure about it. If you want, you can set a clear plan for when he'll pay you back so it feels more manageable.	That's really generous of you—lending $200 to help your brother out is a big trust move. If it helps, you can set a gentle reminder for when he should pay you back so it feels fair for both of you.

Table 17: Blind LLM-judge results against specialized text-only empathetic dialogue baselines on the matched 1,572-example subset.

Model	Win Rate (%)↑	Avg Rank↓	Emotion Fit↑	Humanity ↑	Helpfulness ↑	Naturalness ↑	Overall ↑
EEG-Conditioned LLM (ours)	31.90	2.02	7.91	7.86	6.87	8.20	7.82
Text-Only Baseline	28.63	2.17	7.50	7.54	**7.04**	**8.23**	7.60
EmpDG	0.57	4.46	2.68	2.92	1.88	3.01	2.37
MIME	0.00	4.45	2.70	3.08	1.91	2.69	2.32
Gold-Label Oracle	**38.90**	**1.89**	**8.15**	**8.04**	7.03	8.19	**8.00**

used the author-released outputs and kept only matched cases. The resulting evaluation set contains 1,572 examples and uses the same blind LLM-judge protocol as the main experiments.

Table 17 shows that the EEG-conditioned model outperforms the specialized text-only systems under the same blind evaluation protocol. EmpDG and MIME are useful reference baselines, but they are not matched-backbone controls for isolating the effect of EEG. The matched Text-Only Baseline therefore remains the primary EEG-ablation baseline. In pairwise matched cases, the EEG-conditioned model wins against EmpDG and MIME in 98.1% and 99.2% of cases, respectively.

F.2 EEG Foundation-Encoder Hard-Label Baseline

We also evaluate an EEG foundation-encoder baseline based on frozen NeuroLM Jiang et al. (2025) representations. The baseline trains a linear POS/NEG classifier on top of frozen NeuroLM features, then uses the predicted hard label for explicit label-based LLM control. This directly tests the simplified off-the-shelf foundation-encoder pipeline.

As shown in Table 18, frozen NeuroLM features underperform the task-specific EEGNet encoder on POS/NEG EEG decoding, and the corresponding hard-label control also underperforms in final LLM-judge evaluation. This result does not rule out gains from fully fine-tuned NeuroLM; it tests the simpler frozen-feature pipeline. We do not include EEG-CLIP Camaret Ndir et al. (2025) as a direct off-the-shelf baseline because it aligns TUH clinical EEG with clinical text reports for

Table 18: EEG foundation-encoder hard-label baseline on the 1,597-example evaluation set. EEG Acc/F1 reports POS/NEG EEG decoding accuracy and macro-F1.

Pipeline	EEG Acc/F1	Win%↑	Rank↓	EFit↑	Hum↑	Help↑	Nat↑	Overall↑
EEGNet + projector (ours)	83.02/82.94	**53.79**	**1.54**	**7.59**	**7.56**	**6.66**	**7.93**	**7.52**
EEGNet + hard label	83.02/82.94	24.92	1.66	7.13	7.48	6.48	7.81	7.20
NeuroLM + hard label	65.63/65.26	21.29	1.78	6.84	7.31	6.36	7.71	6.98

pathology-, age-, and gender-style tasks, which is a domain and task mismatch for SEED affective EEG.

G Reproducibility Details

G.1 Computing Environment and Release Plan

All training and offline generation experiments were run on an Ubuntu 22.04.5 LTS workstation with an Intel Core i9-12900KF CPU, 32 GB RAM, and an NVIDIA GeForce RTX 4090 GPU with 24 GB VRAM. The software environment used Python 3.12.4, PyTorch 2.4.1, Transformers 4.57.1, PEFT 0.17.1, NumPy 1.26.4, SciPy 1.13.1, scikit-learn 1.5.1, and pandas 2.2.3.

The online experiment ran on a three-machine setup: the Ubuntu inference workstation, a Windows 7 acquisition/control computer, and a Windows 10 experiment-interface computer. EEG acquisition used a NeuroScan 64-channel wet-electrode cap. We will release all code upon acceptance, including training and evaluation code and preprocessing scripts. Split manifests, generated evaluation files, and anonymized online-study logs will be released upon publication, with all released artifacts subject to source-compatible non-commercial terms. EmpatheticDialogues-derived artifacts will retain attribution to the original dataset and the CC BY-NC 4.0 non-commercial usage condition. The prompts used for data construction and blind judging are already included in Appendices A.3 and D.3. The released code will include a README mapping the main scripts to the paper sections and comments for the newly introduced modeling, preprocessing, evaluation, and online-inference components. The original SEED and EmpatheticDialogues files are not redistributed because they are existing public datasets; users can obtain them from the original dataset providers and then run the released scripts and manifests.

G.2 Data Splits and Preprocessing

For EEG, the final pipeline uses SEED under a fixed cross-subject split: subjects 1–9 for training, subjects 10–12 for development, and subjects 13–15 for testing. Neutral trials are omitted, and the original SEED labels are mapped to the binary affect labels POS and NEG. Raw EEG is segmented into non-overlapping 4-second windows at 200 Hz, giving tensors of shape 62×800. The Stage A preprocessing computes channel-wise z-score statistics from the training data only and applies the same statistics to validation, development, and test windows. These statistics are stored in the Stage A bundle and reused by downstream and online inference code.

For the text side, the source data are derived from EmpatheticDialogues Rashkin et al. (2019). Fine-grained emotions are mapped to POS or NEG, while categories outside this binary setting are discarded. The preprocessing normalizes whitespace and dataset placeholders, groups utterances by conversation, keeps two-speaker conversations, and creates one training example for each assistant turn using the prior dialogue history as input. When official validation and test CSV files are supplied, the code preserves those splits; otherwise, it performs a conversation-level split with seed 42 to avoid turn-level leakage.

The online server follows the offline tensor format. It keeps a rolling EEG buffer, selects a 4-second window spanning video end minus 5 seconds to video end minus 1 second, drops auxiliary/event channels including `HEO`, `VEO`, `M1`, and `M2`, applies a fourth-order 75 Hz Butterworth low-pass filter by default, resamples to 200 Hz, and outputs a 62×800 model tensor.

Table 19: Planned release inventory for data and evaluation artifacts.

Artifact	Source	Released content	Release scope
EmoDial-Binary	EmpatheticDialogues-derived text supervision	POS/NEG dialogue-response JSONL files, split manifests, and preprocessing scripts	Public after publication; CC BY-NC 4.0 terms for source-derived text
EmoDial-Pair	EmpatheticDialogues-derived paired supervision	POS/NEG paired-response JSONL files, construction decisions, and filtering metadata	Public after publication; CC BY-NC 4.0 terms for source-derived text
Pair-Suitable Test Set	EmpatheticDialogues test split	Paired test inputs, model outputs, blind-judge input files, and aggregated judge scores	Public after publication; CC BY-NC 4.0 terms for source-derived text
Offline EEG processing	SEED-derived EEG windows	SEED split manifests, preprocessing scripts, and processed-window generation code; original SEED files excluded	Public after publication
Online user study	Newly collected real-time EEG study	Anonymized raw EEG, trial metadata, participant ratings, model outputs, and derived evaluation files	Public after publication
Reproducibility code	Training, evaluation, and online inference	Script-to-paper README, commented source code, and configuration files for each stage	Public after acceptance

G.3 Randomness, Runs, and Model Selection

All training stages use the fixed seed 42 unless otherwise stated. The code sets Python `random`, NumPy, PyTorch, and CUDA seeds, and the Stage C data loaders use a seeded PyTorch generator. The Stage C pipeline passes the same seed to C0, C1, and evaluation. Final hyperparameters and sweep ranges are reported in Appendix C; Stage A and Stage B are selected by development performance, while Stage C uses the fixed configuration in Table 10.

Reported final model results use one training run for each selected configuration unless a table explicitly states an ablation sweep. The blind LLM-judge evaluation uses three independent judge passes per compared sample and averages the scores across those passes. Candidate identities are hidden from the judge, and candidate order is randomized in the blind evaluation prompt.

G.4 Statistical Tests, Online Study, and Ethics

The main offline comparison uses the dialogue sample as the paired unit after aggregating the three judge passes. It is evaluated with an exact paired sign test on sample-level ranks and paired bootstrap confidence intervals. The online study uses the participant as the paired unit after averaging each participant's 24 trials; it reports two-sided Wilcoxon signed-rank tests and subject-level bootstrap confidence intervals for the main paired comparisons. Questionnaire-style human ratings are reported descriptively.

The online study recruited 18 university students, each completing 24 trials. Participants reported no history of neurological disorders, psychiatric conditions, or substance abuse, provided written informed consent before the study, and received fixed monetary compensation. Released online-study files will be anonymized before publication.