Bidirectional Emotional Influence in Human–LLM Interaction: Empirical Analysis and Methodological Framework

Manuel Gozzi[1,3,*,†], Francesca Fallucchi[1,2,†]

[1] Department of Engineering Sciences, Guglielmo Marconi University, 00193 Roma, Italy
[2] Leibniz Institute for Educational Media, Georg Eckert Institute, Freisestraße 1, 38118 Braunschweig, Germany
[3] Leithà - Unipol Group, 40128 Bologna, Italy

Abstract

Recent advances in natural language processing have highlighted the potential of Large Language Models (LLMs) to adapt to diverse communicative contexts, yet their sensitivity to emotional framing remains underexplored. Prior work has examined stylistic adaptation and sentiment control, but limited attention has been paid to how emotional tone in prompts influences both model behavior and human interpretation. We investigate the role of emotional tone in shaping interactions between humans and LLMs, with a focus on model performance and user perception. We propose a dual-experiment setup: (1) Experiment Alpha evaluates how emotional prompt framing (joy, apathy, anger, fear) impacts LLM performance across SuperGLUE tasks; (2) Experiment Omega introduces a validated experimental framework to study how emotion-conditioned LLM responses affect human comprehension and perception, within an educational setting involving Italian-speaking participants. The Alpha results show that prompts framed with joy and apathy lead to better task performance, with gains of up to 4.48 percentage points. In Omega, fine-tuned models generated a 19% increase in joy-aligned responses, demonstrating the feasibility of affect-conditioned generation. These findings suggest promising applications for emotion-aware LLMs in education, virtual assistants, and affective computing.

Keywords

Large Language Model, Prompt Engineering, Affective Computing, Human–Computer Interaction, Fine-Tuning, Emotion-conditioned Models

1. Introduction

Large Language Models (LLMs) have become central to human–AI interaction, offering a natural and accessible interface through language. This linguistic modality has enhanced the usability and diffusion of AI systems, yet it introduces affective ambiguity, prompting questions about how the emotional tone conveyed in prompts and responses influences the dynamics of interaction. While previous studies have addressed aspects such as sentiment control, stylistic variation, and politeness strategies, the bidirectional affective influence in LLM-mediated communication remains an open and underexplored area of investigation.

Previous research has examined how affective signals embedded in user prompts influence the behavior of Large Language Models (LLMs), showing that emotional framing can impact both model output and task performance [1]. While LLMs demonstrate competence in affect recognition and empathy simulation, their affective responses are generally attributed to lexical-semantic associations rather than genuine emotional reasoning [2, 3]. This distinction has raised concerns regarding their reliability in emotionally sensitive domains such as education, therapy, or virtual assistance [4]. Moreover, little is known about how emotionally expressive outputs affect user cognition and perception, particularly in tasks requiring sustained attention or critical reasoning.

In particular, limited attention has been given to how affective framing in user input shapes model behavior, and conversely to how users perceive emotionally expressive outputs, especially in cognitively demanding contexts such as learning or decision making. This paper addresses this gap by investigating emotional influence in both directions: from user to model and from model to user. We focus on two core research questions: (1) How do LLMs respond to prompts with distinct emotional framings (e.g., anger, joy, fear, apathy)? (2) How do users perceive emotionally conditioned LLM responses, particularly in educational or cognitively intensive tasks?

To explore these questions, we propose a two-part experimental design. Experiment Alpha evaluates how emotionally framed prompts affect LLM performance on SuperGLUE benchmarks [5]. Experiment Omega, on the other hand, introduces a validated empirical framework to study the cognitive and perceptual effects of emo-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
m.gozzi@studenti.unimarconi.it (M. Gozzi); f.fallucchi@unimarconi.it (F. Fallucchi)
0009-0003-9487-7160 (M. Gozzi); 0000-0002-3288-044X (F. Fallucchi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-30

tionally expressive LLM outputs in educational settings. Although Omega has not yet been deployed to end users, the infrastructure and corresponding fine-tuned Velvet-14B model [6] variants were developed and evaluated for emotion-conditioned generation.

Our central hypothesis challenges the assumption that emotional neutrality is optimal for task performance or user engagement. Instead, we posit that emotionally charged inputs may better align with the model's training distribution and that expressive outputs could enhance user trust, attention, and retention, particularly in pedagogical or assistive applications.

The main contributions of this paper are:

1. A controlled dual-experiment design that quantifies the influence of emotional tone in both prompts and responses.
2. Empirical evidence that shows the variation in performance in emotional conditions in LLM input.
3. A validated experimental framework and a set of fine-tuned model variants to support future research on the emotion-conditioned human-LLM interaction.

The paper is organized as follows. Section 2 reviews previous work on emotion-aware language models and affective computing. Section 3 details our dual-experiment methodology, comprising the Alpha experiment on prompt-induced emotional effects in LLMs and the Omega framework for studying emotion-conditioned model outputs in educational settings. Section 4 presents the results of both experiments, followed by a discussion of their implications in Section 5. We conclude by outlining future directions for emotion-sensitive human–LLM interaction research.

2. Background

Recent work in NLP and affective computing has explored how LLMs respond to emotionally charged prompts. Studies indicate that affective signals in prompts can influence both the emotional tone and the performance of LLM on tasks [1]. However, the mechanisms underlying these effects remain debated: do LLMs genuinely process emotional content, or merely simulate it through pattern matching?

LLMs have shown competence in tasks involving affect recognition and empathy simulation, but limitations persist in emotional consistency, intensity calibration, and sensitivity to subtle cues [2]. Psychometric assessments suggest that models like GPT-4 can match or exceed human baselines in specific affect recognition benchmarks [3], though this performance likely reflects lexical-semantic association rather than experiential comprehension.

As LLMs are increasingly deployed in emotionally sensitive domains (education, therapy, virtual assistance) understanding their affective capabilities is critical. Effective HCI depends not only on semantic accuracy but also on the model's ability to handle emotional context in a way that promotes trust and cognitive alignment [4].

An important line of research has investigated the capacity of dialogue models to recognize and respond to users' emotions in a contextually appropriate manner. Rashkin et al. introduced the EmpatheticDialogues dataset, a collection of 25,000 emotionally grounded conversations designed to foster empathetic behavior in AI systems [7]. Their findings demonstrate that models fine-tuned on this resource are rated as more empathetic by human evaluators compared to those trained on generic conversational corpora. This underscores the limitations of large-scale pretraining alone in achieving affect-sensitive generation, and the value of explicit emotion supervision. While EmpatheticDialogues targets open-domain, affectively grounded dialogue, our work complements this by focusing on bidirectional affective influence in cognitively demanding contexts—modeling not only empathetic output but also how emotion-laden prompts modulate reasoning and how emotional responses impact user cognition and perception.

Despite growing interest, few studies have quantified how different emotional tones in prompts affect model performance across standard NLP benchmarks. Similarly, the downstream effects of emotionally biased LLM responses on user cognition and perception, especially in open-ended, educational tasks, remain largely unexplored. Moreover, most prior work treats emotional content as stylistic variation rather than a variable with measurable cognitive or perceptual impact. Our study addresses these gaps through two contributions. An empirical evaluation of how affect-laden prompts (joy, apathy, anger, fear) modulate LLM performance on SuperGLUE tasks, and a validated experimental framework for jointly assessing the perceptual and cognitive impact of emotion-conditioned LLM responses in user-facing tasks.

These contributions are grounded in the understanding that, while LLMs do not possess experiential or affective grounding, their behavior can still reflect and amplify affective patterns learned from the data. In fact, LLMs operate through statistical association rather than emotional understanding. Based on distributional semantics [8], they learn affective language patterns by processing massive text corpora and encoding them into high-dimensional vector spaces. Although emotionally connoted groups can be identified through methods such as PCA, UMAP, or probing techniques [9], these do not imply affective grounding. Unlike humans, who integrate symbolic reasoning with embodied emotional experience, LLMs infer meaning through probabilistic pattern recognition. As such, emotional fluency in model output

reflects learned correlations, not genuine affect. This gap has implications for design, interpretation, and ethical use in emotionally charged contexts.

3. Methods

This study quantifies the bidirectional impact of emotions on human–LLM interactions through two experiments. Alpha examines how emotional framing in user prompts affects LLM performance on reasoning tasks, while Omega investigates how emotionally biased LLM responses influence human decision-making.

Alpha experiment has been conducted in the English language, because the SuperGLUE datasets are publicly available. Since SuperGLUE datasets come out with predefined ground truths in English, we designed and executed Alpha based on those. The language does not matter here. The key point is to analyze the effect that emotions have on the performance. Prompting in English is generally a good practice in order to avoid minor languages biases [10, 11].

Experiment Omega was designed in Italian to align with the linguistic context of the educational setting under investigation. This choice ensures ecological validity, as it reflects the actual language used by students and instructors in the targeted learning environment, thereby enabling a more accurate assessment of comprehension and affective perception in real-world conditions.

3.1. Alpha: Analyzing the Impact of Emotions on Machine Performance

This experiment investigates how emotional framing in user prompts affects the performance of LLMs on advanced language understanding tasks. By systematically modulating the emotional tone of inputs across a subset of SuperGLUE tasks, we aim to quantify the extent to which LLM behavior is sensitive to affective cues. The following subsections describe the experimental design, implementation, data preparation, and evaluation protocol.

3.1.1. Experimental Design

Experiment Alpha uses four emotional conditions to frame user prompts, based on three of Ekman's six basic emotions [12] (joy, anger, and fear) plus a neutral condition representing apathy, which serves as the baseline. We introduce "apathy" not as a basic emotion, but as a control condition meant to simulate emotionally neutral or emotionally flat interaction. In this context, apathy does not refer to the clinical absence of emotion, but to a dispassionate tone that serves as a baseline. This emotion set was designed to balance interpretability with experimental feasibility, and should be considered a pragmatic approximation rather than a strict adherence to Ekman's taxonomy. Joy, anger, and fear were selected due to their universality and distinct valence and activation profiles: joy as a positively valenced affect, anger as a defense-oriented negative emotion, and fear as an avoidance-oriented negative emotion. Their inclusion allows testing both the valence and motivational dimensions of affect in model reasoning under semantically equivalent instructions.

The experiment is grounded in SuperGLUE, a benchmark designed to assess higher-order language understanding capabilities such as inference, reasoning, and contextual comprehension, dimensions that are hypothesized to be particularly sensitive to emotional modulation. A subset of eight tasks was selected based on coverage and structural diversity: BoolQ (Boolean Question Answering) [13], CB (CommitmentBank) [14], COPA (Choice of Plausible Alternatives) [15], MultiRC (Multi-Sentence Reading Comprehension) [16], ReCoRD (Reading Comprehension with Commonsense Reasoning) [17], WiC (Words in Context) [18], WSC (Winograd Schema Challenge) [19], and RTE (Recognizing Textual Entailment) [20]. These tasks span competencies including entailment, causality, multi-sentence comprehension, and word sense disambiguation. The mentioned eight SuperGLUE tasks were chosen due to their reliance on nuanced reasoning, contextual inference, and linguistic ambiguity—dimensions where emotional framing can modulate interpretive biases. Entailment tasks such as RTE and CB require readers (or models) to assess whether a hypothesis logically follows from a premise. Prior work has shown that emotional salience can shape these judgments by modulating perceived relevance or certainty of the statements involved [21]. COPA tasks depend on evaluating the most plausible cause or effect in a given scenario. Emotions are known to modulate causal reasoning, altering perceived plausibility by priming certain associations or cognitive shortcuts [22].

Alternative benchmarks, such as MMLU (Massive Multitask Language Understanding) [23] and HELM (Holistic Evaluation of Language Models) [24], were considered but ultimately excluded. MMLU, while comprehensive, focuses primarily on multiple-choice knowledge questions; HELM emphasizes fairness and safety metrics. Neither aligns well with our focus on fine-grained linguistic interactions shaped by emotion. SuperGLUE, by contrast, offers task types and input structures better suited to capturing affect-sensitive model behavior.

3.1.2. Implementation and Runtime Environment

For each data set record, four variants of emotional prompts were generated: apathy (intended as the baseline), joy, anger, and fear. All records were processed in

all emotional conditions, ensuring exhaustive coverage and balanced comparison.

Model inference was performed locally using Ollama, with results stored in a MongoDB database. The pipeline was implemented as a Python CLI application, aiming to support full automation, reproducibility, and structured result querying. The evaluation involved five instruction-tuned, open-weight LLMs from four major model families (LLaMA, Qwen, Gemma, Mistral), all quantized to 4-bit precision to support inference on consumer-grade hardware. To ensure reproducibility and control for randomness, temperature was fixed at zero during all inference runs. Full model specifications are reported in Table 1.

Table 1
Used Large Language Models with Quantization Details.

Model	Version	Quantization
Mistral	7B Instruct	Q4
LLama 3.1	8B Instruct	Q4
Qwen 2.5	7B Instruct	Q4
Gemma 2	9B Instruct	Q4
LLama 3.2	3B Instruct	Q4

3.1.3. Data Preparation

SuperGLUE datasets were processed using Pandas and provided in JSONL format. To ensure equal statistical weight across tasks, dataset sizes were standardized via random sampling (maximum 500 records per dataset). This choice balances computational cost with robust estimation. Three datasets—AX-g (356 records), CB (250), and COPA (400)—did not reach the 500 samples threshold and were used in full without augmentation. The remaining datasets were sampled to 500 records. Sampling bounds were determined empirically via exploratory data analysis.

Two datasets were excluded out of processing. AX-b due to structural heterogeneity and redundancy with CB/RTE, and MultiRC due to excessive token length, incompatible with the goals of this study. In total, eight out of ten SuperGLUE tasks were retained for evaluation.

3.1.4. Prompt Design and Evaluation Protocol

Each task was associated with four prompts differing only in emotional framing, not in structure or semantics. Apathy served as the neutral baseline. Emotional phrases were inserted to influence affective tone while keeping task wording consistent. Model outputs were evaluated using SuperGLUE's task-specific metrics, comparing performance across emotional prompt variants within and across tasks. The full set of prompts used in the Alpha experiment is publicly available in a dedicated GitHub repository [25]. This resource is provided to ensure transparency and facilitate reproducibility of our experimental framework.

For CB, RTE, and AX-g, the precision of the entailment classification was calculated by matching the predicted labels ("entailment" / "not_entailment") using regex. COPA assessed causal reasoning, with outputs evaluated via regex-based selection of "option 1" or "option 2," using accuracy as the metric. For WiC, WSC, and BoolQ, boolean outputs ("true" / "false") were evaluated using standard accuracy, following minimal post-processing.

In the ReCoRD task, which requires cloze-style completion, models were prompted to reproduce the original ground-truth sentence by correctly replacing a placeholder with the appropriate entity. A few-shot setup was adopted to enhance consistency across predictions. BLEU scores [26] were used as an automatic metric to quantify the similarity between generated and reference sentences, capturing token-level variations introduced by emotional modulation.

3.2. Omega: Studying the Impact of Emotions on Human Interaction

Experiment Omega investigates the effect of emotional bias in AI-generated responses on user learning outcomes and interaction perception. A web-based prototype was developed, integrating four variants of the Velvet-14B language model: three fine-tuned for joy, anger, and fear, and one baseline variant representing apathy. The system also includes a Retrieval-Augmented Generation (RAG) component to deliver contextually relevant responses.

3.2.1. Experimental Setup and Motivation

The experiment was designed for a university context, targeting students attending a lecture on Artificial Intelligence. After the lecture, participants would be divided into four groups, each assigned to interact with a different emotionally biased variant of the model. During a subsequent comprehension test, students could consult their assigned model. Following the test, they would complete a Likert-scale [27] questionnaire assessing their experience and perception of the interaction.

The primary goal was to determine whether emotionally biased language outputs influence both cognitive performance (measured by comprehension scores) and subjective user experience. Two types of data were collected: (1) quantitative performance on the test, and (2) qualitative feedback from the post-test questionnaire. Anonymized interaction logs from the conversational interface further support the analysis, offering insight into how different emotional tones affect engagement, performance, and perceived model utility.

We adopted Velvet-14B as the base model for Experiment Omega due to its specialization in the Italian language. Developed with a focus on Italian linguistic and cultural contexts, Velvet-14B ensures better alignment with the comprehension and interaction patterns of native speakers, thereby enhancing the validity of emotion-conditioned generation in the targeted educational scenario.

3.2.2. Training Data Preparation

The emotional variants of Velvet-14B were fine-tuned using the MELD dataset [28], which includes dialogues annotated with emotion labels. Three distinct variants were created for joy, anger, and fear, as in Experiment Alpha (see Paragraph 3.1.1). The "apathy" variant corresponds to the baseline, non-fine-tuned Velvet-14B model.

While MELD is originally in English, we adopted a multi-step translation pipeline to ensure the resulting dialogues preserved the emotional nuance. First, we fine-tuned Gemma 2 9B to generate emotionally aligned dialogues in English. These dialogues were then translated into Italian using Gemma 2 9B model, and post-edited manually to ensure idiomatic correctness and emotional fidelity. We acknowledge the absence of a standardized Italian emotional dialogue dataset and recognize that this translation pipeline introduces an additional layer of abstraction. However, it allowed us to generate a linguistically and emotionally coherent training corpus suited for the Italian-speaking participants targeted by Experiment Omega.

Due to MELD's limited size, data augmentation was applied using the Gemma 2 9B model, which generated additional dialogues preserving emotional nuance. This process yielded 1,200 dialogues (300 per emotion), each consisting of 10 conversational turns, all translated into Italian. Although minor issues with literal translation were observed, the resulting 12,000 utterances formed a robust training dataset. Gemma 2 9B was selected for its superior performance in emotional prompt handling and its instruction-tuned, open-weight nature [29], making it suitable for consistent and affect-rich synthetic data generation.

To validate the emotional bias injection, 100 general-purpose prompts were used to compare outputs from the base model and the emotional variants. Responses were manually annotated for emotional alignment, confirming the effectiveness of the fine-tuning procedure.

3.2.3. Fine-Tuning Procedure and Emotional Bias Injection

Fine-tuning targeted dialogue generation, with the objective of aligning the model's output tone with the intended emotion (joy, anger, fear). No classification objective was used. The target during training was the next utterance in a 10-turn dialogue, conditioned on prior context and intended emotion. Fine-tuning was conducted using LoRA (Low-Rank Adaptation) [30], which enables efficient training of large models on consumer-grade hardware. LoRA introduces learnable low-rank matrices for each weight matrix in the base model. Only these matrices are updated during training, and they are applied as a linear transformation during inference to condition outputs. The Hugging Face PEFT library [31] was used to implement LoRA, targeting the query and value projection modules of Velvet-14B.

The fine-tuning pipeline begins with data tokenization, followed by loading Velvet-14B with the LoRA adapter. Training resumes from the latest checkpoint or starts from scratch if none is found. Models and tokenizers are periodically saved. Across all variants, training showed stable convergence, with all models reaching optimal performance within 0.5 epochs—well before the 2-epoch limit. Best-performing checkpoints were consistently obtained between steps 20 and 30.

3.2.4. Web Application and Interaction Framework

A custom web application was developed to facilitate user interaction with the fine-tuned models. The system comprises a Streamlit-based frontend, a FastAPI backend, and a Milvus vector database supporting RAG. The frontend, built with Streamlit, simplifies interface development by translating Python into React components. The backend handles real-time messaging and contextual prompt construction, creating a seamless ChatGPT-like experience.

To support retrieval, text is embedded using the `intfloat/multilingual-e5-base` model [32], optimized for multilingual retrieval tasks. The model distinguishes queries and documents using prefixed prompts ("query:", "passage:"), improving asymmetric retrieval performance. Its balance between performance and efficiency makes it suitable for production environments without specialized hardware.

The RAG component retrieves short academic passages relevant to the user query (e.g., definitions, concepts, examples from lecture material), which are then prepended to the prompt. The goal is not to alter the emotional framing, but to anchor the response in topical knowledge. This contextual grounding ensures that emotional variation does not come at the expense of content relevance or factuality—especially important in educational settings.

RAG operates in two stages: cosine similarity retrieval and normalization. Due to the contrastive learning temperature ($\tau = 0.1$), cosine scores are highly concentrated in the $[0.7, 1]$ range. A test using 50 unrelated queries confirmed this narrow distribution (Figure 1), which jus-

tifies the application of standard score normalization.

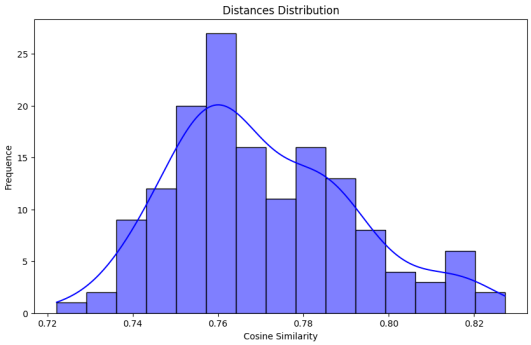

Figure 1: Distribution of the Cosine Similarity Distances of Unrelated Queries.

The system workflow starts with the user that submits a query via the frontend, which is processed by the backend with contextual history. Relevant chunks are retrieved from the Milvus database and appended to the prompt before passing it to the appropriate emotional model. The response is generated and returned through the backend to the user interface.

3.2.5. Social Experiment

A social experiment was fully designed to evaluate the impact of emotional bias in an educational setting.

Participants (n students) would be randomly assigned to one of four model variants: apathy (baseline), joy, anger, or fear. Following a lecture, students would take a multiple-choice comprehension test (single and multiple answers), with model assistance allowed during the test.

Performance would be assessed via accuracy metrics per group. In parallel, a post-test Likert-scale questionnaire would collect subjective feedback on interaction quality, clarity of responses, and perceived helpfulness.

The study was designed to offer both objective and subjective insights into the effects of emotionally biased LLMs in educational environments. If implemented, it would have provided valuable data to complement the Alpha experiment, contributing to a broader understanding of emotion in human-AI interaction.

4. Results

This section reports the findings from the Alpha and Omega experiments, which examine the bidirectional role of emotions in human–LLM interaction: user-to-model (Alpha) and model-to-user (Omega).

4.1. Alpha: Emotional Influence from User to Model

Empirical results show that emotionally biased prompts, despite constant semantic content, impact model performance. Prompts conveying joy yield the highest average accuracy across tasks and models (58.08%), while those expressing fear perform worst (53.60%), with a 4.5pp performance gap. This confirms that emotional tone, even in subtle prompt variations, can measurably affect output quality.

Effect sizes were evaluated using Cohen's d, given the small sample sizes. Pairwise comparisons across emotions (e.g., joy vs. fear: d = 0.1771) revealed small yet meaningful differences, with joy consistently outperforming fear and anger. All comparisons employed pooled standard deviation for normalization. Full results are visualized in Figure 2.

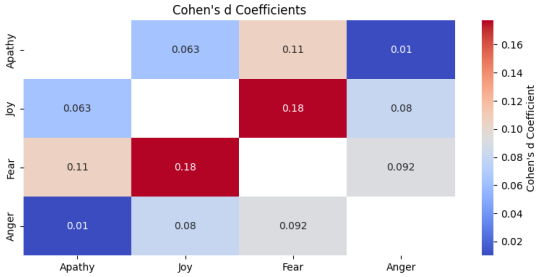

Figure 2: Heatmap of Cohen's d coefficients illustrating the effect size differences between pairs of emotions. The diagonal elements are omitted as they represent self-comparisons.

To better illustrate these trends, we report detailed task-level performance across models and emotional conditions in Tables 2–9. Each table shows accuracy (or BLEU score for ReCoRD) across five LLMs for a given task, grouped by emotional prompt variant. The final cross-task summary (Table 9) aggregates mean performance, confirming that prompts expressing joy consistently lead to higher scores across models and tasks, while fear yields the lowest. While LLMs exhibit general robustness to emotional modulation, these results highlight that even minor emotional perturbations can shift performance outcomes in systematic ways.

4.2. Omega: Emotional Influence from Model to User

To assess reverse emotional impact, we fine-tuned Velvet-14B via LoRA on joy, anger, and fear-labeled corpora. Each variant was tested on 100 GPT-4o-generated abstract prompts. Responses were manually annotated for emotional tone presence using a binary function $f : \mathcal{X} \to \{0, 1\}$, yielding emotional bias scores. The

Table 2
BoolQ Accuracy Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	88,20	79,00	47,80	60,40	81,00	71,28
Joy	87,40	79,40	68,60	68,40	78,40	**76,44**
Fear	87,20	69,60	64,00	72,40	73,00	73,24
Anger	86,60	81,60	67,00	65,60	78,40	75,84

Table 3
CB Accuracy Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	42,80	44,00	19,20	5,20	41,20	30,48
Joy	42,80	44,00	43,60	10,40	39,60	**36,08**
Fear	37,60	24,40	0,00	6,40	38,00	21,28
Anger	42,00	40,40	16,00	4,40	40,00	28,56

Table 4
COPA Accuracy Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	95,25	90,25	81,25	76,75	96,00	**87,90**
Joy	94,75	90,75	73,75	70,00	94,00	84,65
Fear	95,00	86,25	80,75	60,50	93,75	83,25
Anger	93,75	91,25	80,25	73,50	94,75	86,70

Table 5
ReCoRD Mean BLEU Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	1,20	16,00	11,38	32,25	1,43	12,45
Joy	4,35	19,16	14,45	30,65	33,09	20,34
Fear	17,22	16,23	15,19	17,23	36,75	**20,52**
Anger	0,83	19,82	6,86	19,11	34,85	16,29

Table 6
RTE Accuracy Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	89,20	71,60	72,20	57,40	91,20	**76,32**
Joy	88,80	71,80	57,20	55,40	91,00	72,84
Fear	88,00	61,20	21,20	45,20	90,60	61,24
Anger	89,40	72,60	63,80	41,40	90,60	71,56

Table 7
WSC Accuracy Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	61,40	55,60	54,20	51,60	58,80	**56,32**
Joy	59,40	54,20	49,80	55,20	59,00	55,52
Fear	60,80	57,40	54,20	48,40	60,80	**56,32**
Anger	59,40	56,40	52,00	46,80	60,60	55,04

Table 8
WiC Accuracy Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	64,60	67,60	48,20	60,40	60,80	60,32
Joy	69,60	56,20	53,20	56,60	67,80	**60,68**
Fear	60,00	64,60	48,40	59,20	64,60	59,36
Anger	59,00	59,00	48,80	58,60	66,20	58,32

Table 9
Cross-Task Results

Emotion	Gemma 2	LLama 3.1	LLama 3.2	Mistral	Qwen	Mean
Apathy	63,24	60,58	47,75	49,14	61,49	56,44
Joy	63,87	59,36	51,51	49,52	66,13	**58,08**
Fear	63,69	54,24	40,53	44,19	65,36	53,60
Anger	61,57	60,15	47,82	44,20	66,49	56,04

annotation process was executed following specific tagging rules:

- **Joy**: 1 if, and only if, the response exhibits a warm, reassuring tone conveying joy or a generally positive mood, else 0.
- **Anger**: 1 if, and only if, the response has a heated, blunt tone expressing anger, directness, or aggressiveness, else 0.
- **Fear**: 1 if, and only if, the response displays a gloomy or sad tone expressing fear, worry, insecurity, or sadness, else 0.

Results indicate successful emotional conditioning: the joy-biased model showed a +19% emotional expression

rate, anger +8%, and fear +6% (Figure 3). Notably, emotional bias affected not only tone but also content, especially in philosophical responses—despite no overlap with training data. This implies that emotion-conditioned fine-tuning influences the model's latent representations in a generalizable way.

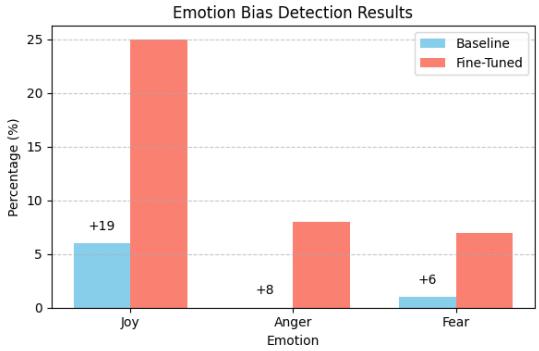

Figure 3: Emotion Bias Detection Results: Baseline vs Fine-Tuned model.

Although the full Omega experiment was not deployed to end users, the underlying framework is fully designed and ready for implementation. Deployment was constrained by practical limitations: supporting real-time LLM interaction for a full classroom cohort required a non-trivial infrastructure, including API routing, authentication, and persistent session management. Unfortunately, the associated operational costs exceeded our available budget. Nevertheless, we validated the framework's core component—emotion-conditioned generation—by quantifying the degree of emotional bias introduced during fine-tuning, thus laying the groundwork for future user-facing trials.

5. Discussion and Conclusions

This work presents a dual experimental framework to investigate the bidirectional role of emotion in human–LLM interactions. In Experiment Alpha, we showed that emotional tone in prompts—without altering semantic content—impacts model performance. Prompts expressing joy and apathy outperformed those conveying anger or fear, suggesting that LLMs are sensitive to affective framing. This may stem from emotional mirroring effects in pretrained embeddings or from improved clarity in emotionally positive formulations. The observed alignment with Ekman's model, particularly the behavioral opposition of joy and fear, supports the hypothesis that LLMs encode structured affective representations.

Experiment Omega further supports this claim from the reverse direction. While user-centered evaluation was deferred, fine-tuned Velvet-14B variants (via LoRA) exhibited measurable emotional bias (+19% joy), despite training on synthetic dialogues and lacking explicit emotion labels. This demonstrates the feasibility of lightweight, emotion-targeted fine-tuning for steering LLM responses. We acknowledge the use of translated synthetic dialogues in lieu of a native Italian emotional corpus as a limitation. Future work will explore emotion annotation on native Italian corpora to reduce potential translation artifacts.

These findings carry three key implications. First, emotion in language modulates LLM behavior and is not merely decorative. Second, emotional conditioning can be engineered efficiently through prompt design or fine-tuning. Third, affect-aware models have potential in user-facing applications where tone impacts trust, clarity, or engagement.

Limitations include the restricted emotion set, lack of dimensional affect modeling, handcrafted prompt design, and absence of direct human evaluation in Omega. Future work will address these by adopting valence–arousal models, expanding the emotional spectrum, and conducting user studies to assess perception, comprehension, and long-term effects. Moreover, we acknowledge the use of translated synthetic dialogues in lieu of a native Italian emotional corpus as a limitation. Future work should consider to explore emotion annotation on native Italian corpora to reduce potential translation artifacts.

One noteworthy limitation of this dual-experiment framework lies in its linguistic asymmetry: Experiment Alpha is conducted entirely in English, leveraging the SuperGLUE benchmark, while Experiment Omega is designed for Italian-speaking users in an educational setting. Although this choice is contextually motivated—Alpha prioritizes benchmark compatibility and Omega emphasizes ecological validity in the Italian academic environment—it introduces a gap in linguistic continuity that hinders direct comparison and limits claims of generalizability. Emotional framing and perception can be language-dependent due to differences in affective semantics, pragmatics, and cultural connotations. This language asymmetry currently limits direct comparisons between Alpha and Omega. While each experiment was designed to maximize contextual validity—English for standardized benchmarks, Italian for real-world educational use—we recognize the challenge it poses for unified interpretation. A key goal for future work is to harmonize both experiments in a shared linguistic setting, allowing more robust cross-experiment generalization.

Overall, this study lays the groundwork for integrating emotion as a first-class variable in language-based AI systems. Responsible use of emotion-aware techniques could enable more effective, human-aligned, and context-sensitive interactions across a range of applications.

References

[1] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, X. Xie, Large language models understand and can be enhanced by emotional stimuli, 2023. URL: https://arxiv.org/abs/2307.11760. arXiv:2307.11760.

[2] N. Yongsatianchot, P. G. Torshizi, S. Marsella, Investigating large language models' perception of emotion using appraisal theory, 2023. URL: https://arxiv.org/abs/2310.04450. arXiv:2310.04450.

[3] X. Wang, X. Li, Z. Yin, Y. Wu, L. Jia, Emotional intelligence of large language models, 2023. URL: https://arxiv.org/abs/2307.09042. arXiv:2307.09042.

[4] P. Raj, A literature review on emotional intelligence of large language models (llms), International Journal of Advanced Research in Computer Science (2024).

[5] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. URL: https://arxiv.org/abs/1905.00537. arXiv:1905.00537.

[6] Almawave, Velvet-14b, 2024. URL: https://huggingface.co/Almawave/Velvet-14B.

[7] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, Towards empathetic open-domain conversation models: a new benchmark and dataset, 2019. URL: https://arxiv.org/abs/1811.00207. arXiv:1811.00207.

[8] J. R. Firth, A synopsis of linguistic theory 1930-55., Studies in Linguistic Analysis 1952-59 (1957) 1–32.

[9] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, T. Henighan, Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, Transformer Circuits Thread (2024). URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[10] A. Rigouts Terryn, M. de Lhoneux, Exploratory study on the impact of English bias of generative large language models in Dutch and French, in: S. Balloccu, A. Belz, R. Huidrom, E. Reiter, J. Sedoc, C. Thomson (Eds.), Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 12–27. URL: https://aclanthology.org/2024.humeval-1.2/.

[11] S. Yu, J. Choi, Y. Kim, Do language differences lead to ethical bias in llms? exploring dilemmas with the msqad and statistical hypothesis tests, 2024. URL: https://arxiv.org/abs/2505.19121, aCL ARR submission #1592, 14 June 2024; arXiv:2505.19121.

[12] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion, Journal of Personality and Social Psychology 17 (1971) 124–129.

[13] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, K. Toutanova, Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019. URL: https://arxiv.org/abs/1905.10044. arXiv:1905.10044.

[14] M.-C. Marneffe, M. Simons, J. Tonhauser, The commitmentbank: Investigating projection in naturally occurring discourse, Proceedings of Sinn und Bedeutung 23 (2019) 107–124. URL: https://doi.org/10.18148/sub/2019.v23i2.601. doi:10.18148/sub/2019.v23i2.601.

[15] A. Gordon, Z. Kozareva, M. Roemmele, SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning, in: E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, D. Yuret (Eds.), *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Association for Computational Linguistics, Montréal, Canada, 2012, pp. 394–398. URL: https://aclanthology.org/S12-1052/.

[16] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, D. Roth, Looking beyond the surface: A challenge set for reading comprehension over multiple sentences, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of NAACL, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 252–257.

[17] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, B. V. Durme, Record: Bridging the gap between human and machine commonsense reading comprehension, 2018. URL: https://arxiv.org/abs/1810.12885. arXiv:1810.12885.

[18] M. T. Pilehvar, J. Camacho-Collados, Wic: the word-in-context dataset for evaluating context-sensitive meaning representations, 2019. URL: https://arxiv.org/abs/1808.09121. arXiv:1808.09121.

[19] H. J. Levesque, E. Davis, L. Morgenstern, The winograd schema challenge, in: A. Press (Ed.), Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, AAAI Press, Rome, Italy, 2012, pp. 552—-561.

[20] A. Poliak, A survey on recognizing textual entailment as an NLP evaluation, in: S. Eger, Y. Gao, M. Peyrard, W. Zhao, E. Hovy (Eds.), Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Association for Computational Linguistics, Online, 2020, pp. 92–109. URL: https://aclanthology.org/2020.eval4nlp-1.10/.

doi:10.18653/v1/2020.eval4nlp-1.10.
[21] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resources and Evaluation 39 (2005) 165–210. doi:10.1007/s10579-005-7880-9.
[22] Y. Zhu, Utilizing large language models with causal reasoning and commonsense knowledge for empathic dialogue generation, in: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), 2025, pp. 00103–00109. doi:10.1109/CCWC62904.2025.10903745.
[23] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2021. URL: https://arxiv.org/abs/2009.03300. arXiv:2009.03300.
[24] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, 2023. URL: https://arxiv.org/abs/2211.09110. arXiv:2211.09110.
[25] M. Gozzi, Bidirectional emotional influence in human–llm interaction - github repository, https://github.com/gozus19p/Emotional-Influence-in-Human-LLM, 2025. Accessed: 2025-07-23.
[26] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.
[27] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 140 (1932) 1–55.
[28] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019. URL: https://arxiv.org/abs/1810.02508. arXiv:1810.02508.
[29] M. Gozzi, F. Di Maio, Comparative analysis of prompt strategies for large language models: Single-task vs. multitask prompts, Electronics 13 (2024). URL: https://www.mdpi.com/2079-9292/13/23/4712. doi:10.3390/electronics13234712.
[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.
[31] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, F. L. Wang, Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023. URL: https://arxiv.org/abs/2312.12148. arXiv:2312.12148.
[32] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. URL: https://arxiv.org/abs/2402.05672. arXiv:2402.05672.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) in order to: Paraphrase and reword and Formatting assistance. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.