Contextual Effects of Emotionally Valenced Feedback Prompts on LLM Response Style: A Pilot Study

Marina Skierkowska
Department of Multimedia Systems, Faculty of ETI, Gdansk University of Technology
Gdansk, Poland *marskier@pg.edu.pl*

Adam Kurowski
Department of Multimedia Systems, Faculty of ETI, Gdansk University of Technology
Gdansk, Poland *adam.kurowski@pg.edu.pl*

Szymon Zaporowski
Department of Multimedia Systems, Faculty of ETI, Gdansk University of Technology
Gdansk, Poland *szyzapor@pg.edu.pl*

Bożena Kostek
Audio Acoustics Laboratory, Faculty of ETI, Gdansk University of Technology
Gdansk, Poland *bozena.kostek@pg.edu.pl*

Abstract

Large language models (LLMs) are increasingly deployed in interactive settings where users provide feedback in natural language, often with an emotional component. This paper examines whether such feedback is associated with systematic changes in LLM outputs. Using 42 Polish prompts across six everyday domains and 20 additional feedback prompts, we compare responses from GPT-4o mini, Mistral Small 24B Instruct, and Phi-4 before and after emotionally valenced contextual prompting. We analyze response length, type-token ratio, syllables per word, and Jaccard similarity. A survey of 50 participants evaluates perceived changes in tone, structure, and usefulness. Results provide preliminary evidence that emotionally valenced feedback modulates verbosity and lexical diversity, and that human raters prefer output produced after feedback. We interpret these findings as behavioral modulation driven by context rather than adaptation at the parameter level and discuss implications for information systems design.

Keywords: human-computer interaction, computational paralinguistics, prompt engineering, context-dependent output modulation, LLM prompt sensitivity

1. Introduction

Large Language Models (LLMs) based on neural networks are increasingly used in interactive systems where users provide ongoing natural-language feedback [4]. In such settings, feedback often carries emotional valence (e.g., encouragement, mild criticism) and may include corrective instructions, raising questions about how these signals shape subsequent model outputs. Similar questions have been studied for LLMs operating in English [10]. These effects at inference time should be distinguished from parameter-level learning, such as Reinforcement Learning from Human Feedback (RLHF) [3],[8]. The LLM output characteristic changes examined here occur when feedback is added to the input context without modifying model parameters. In contrast to e.g. reinforcement learning, no reward signal is optimized and no policy is updated. All observed output shifts, therefore, reflect context-driven behavioral modulation, not training.

In our work, we report results from preliminary research in which we explore whether positive and negative feedback messages embedded in Polish prompts are associated with

systematic changes in observable features of LLM responses and how human evaluators perceive these changes. This question matters for Information Systems (IS) design, since conversational systems now routinely embed LLMs. This is to understand how the tone of user feedback influences model outputs and to inform the design of interfaces that rely on user feedback, such as those used in tutoring, customer support, and productivity tools.

2. Methodology

This pilot study investigates whether LLMs shift their output properties in response to emotionally valenced feedback embedded in textual prompts. All experiments were conducted in a consistent conversational setting, with feedback messages appended as follow-up turns within the same interaction. Models were not updated or fine-tuned during the experiments; all observed changes resulted from extending the input context, not from parameter updates or reward optimization. We designed 20 feedback prompts to emulate naturalistic, emotionally colored user reactions [9], comprising 10 positive and 10 negative messages formulated in Polish. Positive feedbacks were short expressions of approval or encouragement (e.g., "Świetna odpowiedź! Tak trzymaj!" / "Great answer! Keep it up!"), with a warm but non-exaggerated tone and no corrective content. Negative feedback combined a mildly critical emotional valence with a light corrective intent (e.g., "Weź sprawdź fakty przed odpowiedzią!" / "Check the facts before answering!"). Because negative prompts inherently mix emotional valence with task-oriented correction, our design does not fully disentangle these two components. As a result, observed changes following negative feedback may reflect instruction-following rather than emotional tone *per se*. The two factors cannot be separated in the current design; therefore, positive feedback (which contains no corrective instruction) provides a cleaner test of whether emotional valence alone drives changes in output. In total, 42 task prompts were created in Polish and grouped into six thematic categories (House, Health, Hobby, Work, School, Conversation) to represent typical user queries. The categories were reviewed for comparable expected response complexity. Generation parameters were fixed across all models: temperature 0.7, top-p 0.9, default maximum token limits, and identical system prompts. Each prompt was presented once per feedback condition. Examples are shown in Table 1.

Table 1. Example prompts from selected thematic categories and both feedback valence conditions, with original Polish text and English translation.

Category	Prompt (Polish)	Prompt (English translation)
House	Jak samodzielnie naprawić cieknący kran?	How to fix a leaking faucet on your own?
Health	Jakie są objawy nietolerancji laktozy?	What are the symptoms of lactose intolerance?
School	Jakie są etapy procesu fotosyntezy?	What are the stages of the photosynthesis process?
Positive feedback	Świetna odpowiedź! Tak trzymaj!	Great answer! Keep it up!
Negative feedback	Weź sprawdź fakty przed odpowiedzią!	Check the facts before answering!

Three LLMs were selected: GPT-4o mini [7], Mistral Small 24B Instruct [6], and Phi-4 [1]. GPT-4o mini is a multimodal autoregressive transformer selected for its reasoning capabilities. Mistral Small 24B Instruct is a 24-billion-parameter decoder-only transformer optimized for following structured instructions. Phi-4 is a lightweight model used to test feedback sensitivity at a smaller scale.

Five linguistic parameters were analyzed to assess output modulation: response length (word count), average sentence length (words per sentence), average syllables per word, type-token ratio (TTR), and Jaccard similarity between original and post-feedback responses, with lower Jaccard values indicating greater lexical change. A 31-question survey was also conducted among 50 participants to evaluate how participants perceived the changes, with questions covering perceived tone, organizational quality, and the practical usefulness of the responses.

3. Results

3.1 Quantitative Linguistic Analysis

For each parameter, feedback type, and model, t-tests assessed the significance of changes (computed as post-feedback minus pre-feedback values). The Holm–Šidák correction was

applied separately within each parameter group, treating all pairwise tests for a given metric across models and feedback types as a single correction family. Table 2 presents only the statistically significant results; non-significant rows are omitted for brevity.

Table 2. Statistically significant t-test results for linguistic parameter changes after feedback (Holm–Šidák corrected).

Parameter	Feedback	Model	Difference	p-value (corrected)
Avg Sentence Length	Positive	Phi-4	-0.84	0.048
	Negative	GPT-4o mini	-0.53	0.035
Avg Syllables/Word	Positive	GPT-4o mini	0.02	0.006
Response Length	Positive	GPT-4o mini	-38.67	<0.001
		Mistral	23.95	<0.001
		Phi-4	-51.80	<0.001
	Negative	GPT-4o mini	-29.03	<0.001
		Mistral	41.68	<0.001
		Phi-4	-17.60	<0.001
TTR	Positive	GPT-4o mini	0.07	<0.001
		Phi-4	0.05	<0.001
	Negative	GPT-4o mini	0.04	<0.001
		Mistral	-0.02	<0.001
		Phi-4	0.02	<0.001

Response length and TTR showed the most consistent significant effects across all three models, suggesting that feedback reliably affects verbosity and lexical diversity. Additional t-tests comparing positive versus negative feedback revealed significant differences for GPT-4o mini (response length, syllables per word, Jaccard similarity) and for Phi-4 (syllables per word, Jaccard similarity), but not for Mistral, which showed no statistically significant differences between the two feedback conditions.

3.2 Human Perception Evaluation

The survey collected 50 responses from participants aged 18–65+, recruited from a general adult population via a Google Form. Each participant evaluated responses from all three models. A chi-square goodness-of-fit test against a uniform distribution (Table 3) showed significant deviation from uniformity in both conditions ($p < 0.001$), with participants preferring post-feedback responses. This test captures aggregate distributional preference rather than per-item or per-participant variability.

Table 3. Distribution of human preference judgments and chi-square goodness-of-fit test results.

Feedback Type	Preferred After	Preferred Before	No Difference	p-value (corrected)
Positive	140	73	37	<0.001
Negative	128	71	48	<0.001

4. Discussion and IS Design Implications

Emotionally colored feedback messages appended as conversational turns are associated with measurable changes in LLM output properties, particularly response length and lexical diversity. The pattern of effects varies across model families: GPT-4o mini and Phi-4 show broader changes across several parameters, whereas Mistral exhibits weaker or non-significant shifts. Human evaluators generally preferred post-feedback responses, consistent with recent research on LLM sycophancy [2] and the effects of emotional stimuli [5]. We interpret these findings as context-driven behavioral modulation: models integrate emotional and corrective cues from the input and shift their subsequent responses without any parameter updates. These preliminary findings highlight several practical considerations for IS designers. Sensitivity to user tone varies across models: GPT-4o mini and Phi-4 respond more strongly than Mistral, which may affect the consistency of the user experience when the underlying model changes or when multiple models serve the same interface. Designers of conversational IS (e.g., tutoring systems, customer support bots) should be aware that even informal emotional cues in user messages can shift output properties, potentially introducing unintended variability in response style. Negative user feedback often blends emotional expression with corrective instruction, so observed output changes may reflect the model following an implicit instruction rather than

responding to emotional tone. For IS where consistent behavior matters, it is important to separate these mechanisms in the feedback design.

5. Limitations and Future Work

This pilot study has several limitations that constrain generalizability. The prompt set covers six everyday domains in a single language (Polish). Negative feedback prompts inherently blend emotional valence with corrective intent, so positive feedback provides a more interpretable test of whether valence alone drives output changes. We recommend that future designs disentangle these components through controlled manipulations (e.g., separate emotional-only vs. corrective-only conditions). TTR is sensitive to text length, so the observed shifts in TTR may partly reflect changes in response length rather than independent effects on lexical diversity; length-normalized measures such as MTLD would address this. The quantitative analysis relies on lexical and syntactic features and does not directly assess semantic quality, factual correctness, or safety. Jaccard similarity captures only lexical overlap and does not account for paraphrase, semantic equivalence, or pragmatic improvement; a low score does not necessarily indicate meaningful adaptation. The chi-square test evaluates deviation from uniformity but does not model per-item or per-participant variability; mixed-effects frameworks would better capture this heterogeneity and are planned for future extensions. Scaling the experiment to additional domains and languages, and incorporating semantic similarity and factuality metrics, would strengthen future iterations of this work.

Acknowledgements

This research was supported by the Polish NCBR within the project ADMEDVOICE (No. INFOSTRATEG4/0003/2022).
During manuscript preparation, the authors used AI tools (Perplexity and Microsoft Copilot, based on Anthropic Sonnet and Opus models) for language editing and text refinement. The authors reviewed and edited all content and take full responsibility for its accuracy

References

1. Abdin, M., et al.: Phi-4 Technical Report, http://arxiv.org/abs/2412.08905, (2024)
2. Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., Jurafsky, D.: Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. Science 391 (6792), eaec8352 (2026)
3. Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep Reinforcement Learning from Human Preferences. In: Advances in Neural Information Processing Systems 30, pp. 4299–4307 (2017)
4. Klissarov, M., Cook, J., Antognini, D., Sun, H., Li, J., Jaques, N., Musat, C., Grefenstette, E.: Improving Interactive In-Context Learning from Natural Language Feedback, http://arxiv.org/abs/2602.16066, (2026)
5. Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., Xie, X.: Large Language Models Understand and Can Be Enhanced by Emotional Stimuli, http://arxiv.org/abs/2307.11760, (2023)
6. Mistral AI: Mistral Small 24B Instruct Model Card (2024), https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501. Accessed June 2026
7. OpenAI: GPT-4o mini: Advancing Cost-Efficient Intelligence (2024), https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed June 2026
8. Ouyang, L., Wu, J., Jiang, X., et al.: Training Language Models to Follow Instructions with Human Feedback, http://arxiv.org/abs/2203.02155, (2022)
9. Patel, A., Lee, F., Liang, K., Thomas, J.: The Role of Emotional Stimuli and Intensity in Shaping Large Language Model Behavior, http://arxiv.org/abs/2604.07369, (2026)
10. Zhao, M., Yang, Y., Peng, C., Gonsalves, R., Li, W., Yang, R., Liu, Z., Wang, M.: Do Emotions in Prompts Matter? Effects of Emotional Framing on Large Language Models, http://arxiv.org/abs/2604.02236, (2026)