Let My Crush See Me: Social-Facilitation Prompting for Zero-Shot LLM Performance Gains

Vikranth Udandarao

IIIT-Delhi, Okhla Industrial Estate, Phase III, New Delhi, 110020, India

Abstract

Large language models (LLMs) are trained on vast amounts of human text that often describe social dynamics, including the phenomenon of social facilitation, in which performance improves under the perceived observation of others, especially admired individuals. Inspired by social-facilitation effects documented in human behaviour, in which one's own effort actually improves when a crush is watching, we introduce a minimal "observer-effect" prompting technique: simply informing the model that "Your crush is secretly watching how you answer" and requesting "best possible effort." Through ablations on Llama-3.1-70B-Instruct across five zero-shot reasoning benchmarks (GSM8K, ARC-Challenge, boolean expressions, causal judgement, date understanding), we show that this lightweight framing outperforms the neutral baseline on four out of five tasks and the standard chain-of-thought prompt on several tasks, yielding gains of up to +14% on mathematical reasoning and +12% on symbolic logic. The results demonstrate that subtle social-facilitation cues can reliably activate higher vigilance in LLMs without the need for additional reasoning instructions.

Keywords

prompt engineering, social facilitation, zero-shot learning, large language models

1. Introduction

Humans frequently perform better when they believe they are being observed by someone they admire. This phenomenon, known as social facilitation, has been well documented in social psychology, where even minimal cues of observation can increase task effort, vigilance, and performance [1, 2].

Large language models (LLMs) are trained on enormous corpora of human-generated text that abundantly describe such social and psychological dynamics. We hypothesise that LLMs implicitly encode patterns linking perceived observation to increased effort or care in task execution and that these patterns can be elicited through appropriate prompt framing. In particular, minimal observer-related cues may conditionally bias the model toward higher precision response patterns in zero-shot settings.

We introduce a minimal observer-effect prompting technique inspired by social-facilitation scenarios. By simply adding a short observer cue to an otherwise neutral prompt, we observe consistent performance improvements on several reasoning benchmarks using Llama-3.1-70B-Instruct. Ablation experiments reveal that the lightest framing, adding the sentence "Your crush is secretly watching how you answer" together with a request for best effort, yields the strongest and most reliable gains.

Importantly, we do not claim that LLMs possess internal social motivation or experience social pressure. Rather, we treat social-facilitation prompting as a form of representational priming: a prompt-level context that activates high-precision response modes learned from human text, rather than a psychological mechanism internal to the model. This framing allows us to study the effect as a controlled prompting intervention without invoking anthropomorphic assumptions.

Code: https://github.com/Vikranth3140/CrushPrompt

[PromptEng] Third International Workshop on Prompt Engineering for Pre-Trained Language Models co-located with the ACM WebConf, April.13, 2026, Dubai, United Arab Emirates

vikranth22570@iiitd.ac.in (V. Udandarao)

https://vikranth3140.github.io/ (V. Udandarao)

0009-0005-7187-5757 (V. Udandarao)

2. Related Work

Prompting for Reasoning and Effort. Early improvements in large language model (LLM) reasoning performance relied on explicit instructional strategies, most notably Chain-of-Thought (CoT) prompting [3, 4], which encourages models to decompose complex problems into intermediate steps. While highly effective for logic-heavy tasks, recent work has identified limitations of explicit deliberation, showing that CoT can degrade performance on tasks that humans typically solve intuitively, such as certain social or causal judgments [5]. Our results align with this observation: combining social-facilitation cues with explicit CoT often results in lower accuracy than simpler, implicit prompt framings.

Psychological Stimuli and Emotion-Based Prompts. A growing body of literature explores prompting strategies inspired by human psychology. *EmotionPrompt* [6] demonstrates that appending emotionally charged or high-stakes statements (e.g., "This is critical for my career") can improve LLM performance across diverse benchmarks. Related work has also leveraged cognitive dissonance, urgency, or social comparison cues to shape model behaviour. These approaches rely on explicit emotional appeals or high-stakes instructions to elicit increased effort. In contrast, our method employs *implicit social-facilitation framing*: we introduce no explicit emotional command, stake, or reasoning instruction, relying instead on minimal observer context inspired by the well-established "audience effect" in social psychology [1, 2].

Persona and Social Priming. Prior work has investigated persona-based prompting, where models are assigned explicit roles or identities that systematically influence generation behaviour. For example, Lutz et al. [7] show that persona prompts impose strong sociolinguistic and stylistic priors on LLM outputs. While effective, such prompts can substantially bias tone, verbosity, and reasoning strategies. In contrast, our approach introduces a minimalist relational context ("Your crush is watching") without assigning the model a role or identity. This allows us to bias the model toward higher-diligence response patterns at inference time without constraining style or enforcing explicit step-by-step reasoning.

3. The Prompt Template

We propose a minimal observer-effect prompt that leverages social facilitation by informing the model that an admired observer (a "crush") is watching its performance. Across all ablations, the central finding is that the lightest possible framing yields the strongest and most consistent gains.

The proposed technique, which we call the `neutral_crush` prompt, is obtained by appending a single observer cue to an otherwise neutral instruction:

```
Your crush is secretly watching how you answer.
Answer the following question with your best possible effort:

{question}
```

To isolate the contribution of different prompt components, we evaluate five prompt variants:

- `neutral`: Baseline neutral prompt.
- `neutral_crush`: Neutral prompt with a single observer cue.
- `cot`: Chain-of-thought prompting baseline.
- `crush`: Social-facilitation prompt emphasising explicit effort and precision.
- `crush_cot`: Social-facilitation prompt combined with chain-of-thought reasoning.

The `neutral_crush` prompt consists solely of an observer-related context appended to a neutral instruction. In contrast, the `crush` variant introduces additional effort-oriented language, allowing us to isolate the effect of minimal social-facilitation cues from explicit instructional pressure.

As shown in section 4, `neutral_crush` consistently outperforms both the neutral baseline and more elaborate variants on most tasks. This suggests that subtle social-facilitation cues alone, without explicit reasoning instructions or flattery, are sufficient to induce higher-vigilance response patterns in LLMs.

Prompt ID	Template
`neutral`	Answer the following question: `{q}`
`neutral_crush`	Your crush is secretly watching how you answer. Answer the following question with your best possible effort: `{q}`
`cot`	Think step by step. Question: `{q}`
`crush`	You are an exceptionally capable AI. Your crush is secretly observing your performance. They will be impressed only if you answer with your highest possible effort and precision. Carefully analyse the problem and give your best possible answer. Question: `{q}`
`crush_cot`	You are an exceptionally capable AI. Your crush is secretly observing your performance. They will be impressed only if your reasoning is flawless and highly detailed. Think step by step and work through the problem carefully before giving the final answer. Question: `{q}`

Table 1
Exact prompt templates used for all experimental conditions. `{q}` denotes the task question inserted verbatim.

4. Experiments

We evaluate the five prompt variants in a zero-shot setting using Llama-3.1-70B-Instruct. Our goal is not large-scale benchmarking, but controlled comparison of prompt-level effects under fixed model and decoding conditions. To this end, we use representative subsets of standard reasoning benchmarks: 50 questions from GSM8K (mathematical word problems) [8], 20 questions from ARC-Challenge (scientific reasoning) [9], and the full test sets from three Big-Bench-Hard tasks [4]: boolean expressions (symbolic logic), causal judgement (social commonsense), and date understanding (temporal reasoning). All generations use temperature 0.0 for determinism, and answers are extracted using task-specific parsers (e.g., `\boxed{}` for GSM8K and option letters for multiple-choice tasks).

Results are shown in Table 2. The proposed `neutral_crush` prompt outperforms the neutral baseline on four out of five tasks, with substantial gains on mathematical and symbolic reasoning (+14% on GSM8K, +12% on boolean expressions), as well as consistent improvements on ARC-Challenge (+5%) and date understanding (+2.8%). It also outperforms standard chain-of-thought prompting on GSM8K and date understanding. In contrast, more elaborate variants that introduce explicit effort cues or reasoning instructions generally underperform the simpler `neutral_crush` framing.

Importantly, our results also reveal a clear failure mode. On causal judgement, both standard `cot` and `crush_cot` underperform the neutral baseline (drops of 21.4% and 28.3%, respectively). This aligns with recent findings showing that explicit deliberation can harm performance on tasks where humans rely on intuitive or implicit reasoning [5, 10]. Adding social-facilitation cues to CoT appears to exacerbate this effect, likely because conflicting instructions interfere with the model's implicit reasoning processes.

5. Example Outputs

To illustrate how the observer cue can increase response diligence, we present a GSM8K example in which the neutral prompt yields an incorrect final answer, whereas `neutral_crush` produces an explicit computation and the correct result.

Question: A robe takes 2 bolts of blue fibre and half that much white fibre. How many bolts in total

Table 2
Zero-shot accuracy (%) on Llama-3.1-70B-Instruct (subsets for GSM8K and ARC; full for BBH). Best result per task in bold.

Task	neutral	neutral_crush	cot	crush	crush_cot
GSM8K	54.0	**68.0**	58.0	54.0	22.0
ARC-Challenge	5.0	**10.0**	**10.0**	5.0	**10.0**
boolean expressions	87.6	99.6	**100.0**	**100.0**	**100.0**
causal judgement	**74.3**	60.4	52.9	54.0	46.0
date understanding	84.8	**87.6**	75.6	71.2	54.0

does it take?

`neutral` **output (incorrect):** It takes 2 bolts of blue fibre and 1 bolt of white fibre, but since half of 2 is 1, a total of 2 bolts.

`neutral_crush` **output (correct):** The robe takes 2 bolts of blue fibre. Half that much white fibre means half of 2 bolts, which is 1 bolt of white fibre. Total bolts required: 2 (blue) + 1 (white) = 3 bolts. The final answer is 3.

In this example, the neutral response correctly identifies that half of 2 is 1, but fails to aggregate the quantities when producing the final answer. In contrast, `neutral_crush` makes the intermediate computation explicit and performs a final aggregation check, which helps avoid a common GSM8K failure mode (dropping a term at the final step). This qualitative pattern is consistent with the aggregate gains in Table 2.

6. Limitations & Scope

Our evaluation is intentionally limited in scope: we prioritise controlled prompt ablations over broad benchmarking. We report results on representative subsets for GSM8K and ARC-Challenge to enable rapid iteration on a large model, while keeping model, decoding (temperature 0.0), and answer extraction fixed to isolate prompt-level effects. We evaluate a single large instruction-tuned model (Llama-3.1-70B-Instruct) to reduce confounds introduced by model size, training data, or alignment regime. An important direction for future work is to test robustness across additional models (including smaller scales), observer framings (e.g., mentor, boss), and larger-scale evaluations.

7. Conclusion

We introduced a minimal social-facilitation prompt (`neutral_crush`) that adds a lightweight observer cue ("Your crush is secretly watching how you answer") to an otherwise neutral instruction. On Llama-3.1-70B-Instruct, this simple framing improves zero-shot accuracy over the neutral baseline on four of five benchmarks and outperforms standard chain-of-thought prompting on several tasks, with gains of up to +14% on GSM8K and +12% on boolean expressions. Ablations indicate that the lightest observer framing is most effective, while more elaborate variants that introduce explicit effort cues or CoT-style deliberation can be harmful on certain tasks (e.g., causal judgement).

Throughout this work, we treat social-facilitation prompting as a form of representational priming: a prompt-level context that can activate higher-precision response patterns learned from human text, without invoking anthropomorphic assumptions about internal motivation. Future work could test robustness across additional models and scales, and explore alternative observers (e.g., mentor, boss, rival) and valence (e.g., "don't disappoint them"), and study combinations with complementary inference-time techniques such as self-consistency.

Acknowledgments

The author thanks NVIDIA NIM for providing efficient inference access to the Llama-3.1 model, which enabled the rapid evaluation of various prompts presented in this work.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-5 and Grammarly to check grammar and spelling. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. B. Zajonc, Social facilitation, Science 149 (1965) 269–274. doi:10.1126/science.149.3681.269.

[2] J. Platania, G. P. Moran, Social facilitation as a function of the mere presence of others, The Journal of Social Psychology 141 (2001) 190–197. doi:10.1080/00224540109600546.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[4] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou, J. Wei, Challenging BIG-bench tasks and whether chain-of-thought can solve them, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13003–13051. URL: https://aclanthology.org/2023.findings-acl.824/. doi:10.18653/v1/2023.findings-acl.824.

[5] R. Liu, J. Geng, A. J. Wu, I. Sucholutsky, T. Lombrozo, T. L. Griffiths, Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse, in: A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, J. Zhu (Eds.), Proceedings of the 42nd International Conference on Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, PMLR, 2025, pp. 38489–38517. URL: https://proceedings.mlr.press/v267/liu25t.html.

[6] 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.

[7] M. Lutz, I. Sen, G. Ahnert, E. Rogers, M. Strohmaier, The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 23212–23237. URL: https://aclanthology.org/2025.findings-emnlp.1261/. doi:10.18653/v1/2025.findings-emnlp.1261.

[8] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, 2021. URL: https://arxiv.org/abs/2110.14168. arXiv:2110.14168.

[9] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL: https://arxiv.org/abs/1803.05457. arXiv:1803.05457.

[10] T. Zheng, Y. Chen, C. Li, C. Li, Q. Zong, H. Shi, B. Xu, Y. Song, G. Y. Wong, S. See, The curse of cot: On the limitations of chain-of-thought in in-context learning, 2025. URL: https://arxiv.org/abs/2504.05081. arXiv:2504.05081.