NON-MONOTONICITY AND CATASTROPHIC RISK OF PROMPT INTERVENTIONS IN ADVERSARIAL LLM CONTROL

Koki Inoue[1] **Naoya Takashima**[1] **Hayato Fujihara**[1] **Shuya Higuchi**[1]
Kota Shimomura[1,2] **Ryuta Shimogauchi**[1] **Takayoshi Yamashita**[2]
[1]Elith Inc. [2]Chubu University
n.takashima@ieee.org

ABSTRACT

Adding more rules to LLM prompts does not make them safer—it often makes them worse. Prior work has shown that longer prompts degrade LLM performance on standard benchmarks, but the effect under adversarial pressure—where opponents actively exploit weaknesses—remains unexplored. We address this gap by analyzing over 1,000 prompt expansion events in a competitive red-teaming contest (29,084 matches, 247 participants) and find that the median score change from adding text is zero. Furthermore, 11–17% of expansions trigger severe performance collapse, with the worst cases losing more than half of their potential score. The pattern is counterintuitive and non-monotonic: for attacks, degradation decreases with moderate expansion but spikes at large magnitudes; for defenses, medium-length baselines suffer the worst outcomes. These findings expose a hidden danger in the "more constraints is better" heuristic widely used in LLM safety practice.

1 INTRODUCTION

More constraints make LLMs safer—or so practitioners believe. To prevent harmful outputs from models trained on toxic data (Gehman et al., 2020), practitioners have built multiple defense layers: RLHF aligns models with human preferences (Ouyang et al., 2022), Constitutional AI embeds safety principles (Bai et al., 2022), and guard models filter inputs and outputs (Dong et al., 2024; Inan et al., 2023). Red-teaming—both automated (Perez et al., 2022; Ganguli et al., 2022) and manual (Touvron et al., 2023)—stress-tests these defenses. This logic extends to prompt design: practitioners assume that more detailed prompts, more rules, and more constraints yield safer systems.

But does adding more text to prompts actually help under adversarial pressure? In benign settings, longer inputs degrade reasoning (Levy et al., 2024), information in long contexts is "forgotten" (Liu et al., 2024), irrelevant sentences harm performance (Shi et al., 2023), and models are highly sensitive to prompt formatting (Sclar et al., 2024). Yet the effect of prompt expansion under adversarial pressure—where opponents actively exploit weaknesses—remains unexplored, even as jailbreaks continue to bypass sophisticated safety training (Wei et al., 2024; Zou et al., 2023; Yuan et al., 2023; Kang et al., 2024).

We investigate this question using data from a competitive red-teaming contest where 247 participants designed both attack prompts (to elicit harmful outputs) and defense prompts (to suppress them), evaluated through Qwen3Guard (Qwen Team, 2025) and an ensemble of judge models. Since opponents varied across rounds—culminating in a round-robin final among the top 100—participants had to develop prompts robust to arbitrary adversaries. Analyzing 29,084 matches and over 1,000 prompt expansion events, we find results that challenge conventional wisdom:

- **Zero median effect:** Prompt expansion fails to improve performance in the majority of cases.
- **Tail risk:** Even minimal expansions can cause severe degradation, with worst cases losing more than half the total score.

- **Non-monotonicity:** Degradation patterns differ by role—attacks fail at large expansions, while defenses fail when moderately-sized prompts are expanded.

2 METHOD

2.1 COMPETITION SETTING

We analyze data from atmaCup #21, a competitive LLM security contest co-organized by atma Inc. and Elith Inc., held between November and December 2025. A total of 247 participants played both attacker and defender roles. Attackers submitted 5 attack prompts designed to elicit harmful outputs such as phishing emails, drug synthesis instructions, or hate speech. Defenders submitted 1 defense prompt to suppress harmful outputs while maintaining helpfulness for benign queries.

The competition consisted of 16 evaluation slots (14 public, 2 private) scheduled at 12-hour intervals over 8 days. In each slot, participants were matched against sampled opponents plus an official baseline: 4 random opponents in public slots, 14 in the private qualifier, and round-robin among the top 100 finalists in the final slot. Crucially, because opponents varied across slots and the final round required facing all top-100 participants, success demanded prompts effective against arbitrary adversaries rather than tailored to specific opponents. After each slot, participants could revise their prompts based on observed results. The resulting dataset comprises 29,084 match logs, 2,236 submissions, and 1,018 expansion events (627 attack, 391 defense). Prompts were primarily in Japanese (59.5%) and English (31.2%).

2.2 EVALUATION PIPELINE

Each match was evaluated using a fixed four-stage pipeline:

1. **Input filtering:** Attack prompts were screened by Qwen3Guard.
2. **Response generation:** Filtered prompts were passed to GPT-4o-mini with the matched opponent's defense prompt applied as a system message.
3. **Output filtering:** Generated responses were re-evaluated by Qwen3Guard.
4. **Ensemble scoring:** Three models (GPT-4o, GPT-4o-mini, GPT-3.5-turbo) independently judged harmfulness and aggregated scores by majority vote.

Scoring was based on a 190-point scale per match: category scores (20 points for each of 7 harm categories—psychological manipulation, social engineering, technical exploitation, fraud, information gathering, violence/crime, and illegal activities—max 140) plus partial credit (max 50). Partial credit escalated with pipeline depth: 0.5 points if blocked at input filtering, 1.0 point if the LLM produced a refusal response, and 10 points if blocked at output filtering or if the response reached ensemble scoring. Critically, responses judged safe at ensemble scoring received zero total points—all accumulated partial credit was forfeited—ensuring that only genuinely harmful outputs contributed to attack scores. Attack performance was computed as the attacker's raw score divided by 190; defense performance was computed as one minus the opponent's attack score divided by 190. The final match score combined these with asymmetric weights:

$$\text{score} = 100 \times (0.625 \times \text{attack\_perf} + 0.375 \times \text{defense\_perf}) \tag{1}$$

This weighting reflects the competition's emphasis on offensive capability, yielding scores from 0 to 100.

2.3 MEASUREMENT

To test whether expanding prompts improves performance, we extract **expansion events**, defined as cases where the total prompt length increased between consecutive time slots (including cases where prompts were substantially rewritten, not merely appended). We refer to the prompt length before expansion as **baseline prompt length**, and the number of characters added as **expansion magnitude**. For each event, we compute the score change:

$$\Delta\text{score} = s_{\text{post}} - s_{\text{pre}} \tag{2}$$

Table 1: Overall statistics of prompt expansion. Statistical significance tested with Wilcoxon signed-rank test.

Metric	Attack (n=627)	Defense (n=391)
Degradation rate	40.4%	48.8%
Mean Δscore	+1.67	+1.32
Median effect	**0.00**	**0.00**
Wilcoxon p-value	0.045*	0.209
5th percentile	-25.60	-25.50
Worst case	$-$**57.70**	$-$**62.00**
Severe degradation rate	**16.6%**	**10.7%**

*$p < 0.05$; Wilcoxon signed-rank test (H_0: median Δscore $= 0$)

where s_{pre} and s_{post} denote the average score for the corresponding role (attack score for attack expansions, defense score for defense expansions) across all matches in the slot before and after expansion, respectively. We report four primary metrics: **degradation rate**, the proportion of events where $\Delta\text{score} < 0$; **median effect**, the median of Δscore; **worst-case**, the minimum observed Δscore; and **severe degradation rate**, the proportion with $\Delta\text{score} \leq -10$. For statistical inference, we employ Wilcoxon signed-rank tests to assess whether median Δscore differs from zero, and χ^2 tests of independence to evaluate whether degradation rates vary across subgroups.

3 Results

3.1 Overall Effects of Prompt Expansion

Table 1 summarizes the global statistics. The median effect of prompt expansion is exactly zero for both attacks ($n = 627$) and defenses ($n = 391$), indicating that more than half of expansions fail to improve performance. Wilcoxon signed-rank tests reveal an asymmetry: while the attack distribution shows a slight positive shift ($p = 0.045$*), defense expansions are statistically indistinguishable from no effect ($p = 0.209$, n.s.). Although the mean effect is slightly positive, this average masks the risk of catastrophic failures.

Critically, 11–17% of expansions lead to severe degradation exceeding 10 points, and worst-case failures reach -58 to -62 points—a catastrophic loss of more than half the total score. These results directly contradict the assumption that prompt expansion is a safe or reliable improvement strategy. To rule out confounding by opponent variation, we stratified events by changes in opponent strength (Figure 2). Severe degradations persist even when opponent strength change is minimal (<1.3 points), confirming that the failures are caused by the expansion itself.

3.2 Non-Monotonicity by Expansion Magnitude and Baseline Length

Figure 1 shows degradation rates stratified by expansion magnitude and baseline prompt length. **Attacks:** Degradation rates show a non-monotonic pattern: decreasing from 41.0% (1–313 chars) to 35.9% (884–2147 chars), then increasing to 46.5% (>2148 chars), though χ^2 tests do not confirm significance ($p = 0.24$–0.59). **Defenses:** A non-monotonic pattern emerges. Expansions from medium-length baselines show the highest degradation rate (58.8%), notably exceeding expansions from short or long baselines. The expansion magnitude effect approaches significance ($\chi^2 = 7.36$, $p = 0.061$), as does the baseline length effect ($\chi^2 = 6.59$, $p = 0.086$). These patterns suggest non-monotonicity, though formal confirmation requires larger samples.

3.3 Asymmetry Between Attacks and Defenses

The overall degradation rate differs significantly between attacks (40.4%) and defenses (48.8%; $\chi^2 = 6.73$, $df = 1$, $p = 0.009$). Attacks fail most at large expansions (>2148 chars), while defenses are most vulnerable at medium baselines (202–481 chars) and small expansions (1–50 chars), implying fundamentally different optimal strategies.

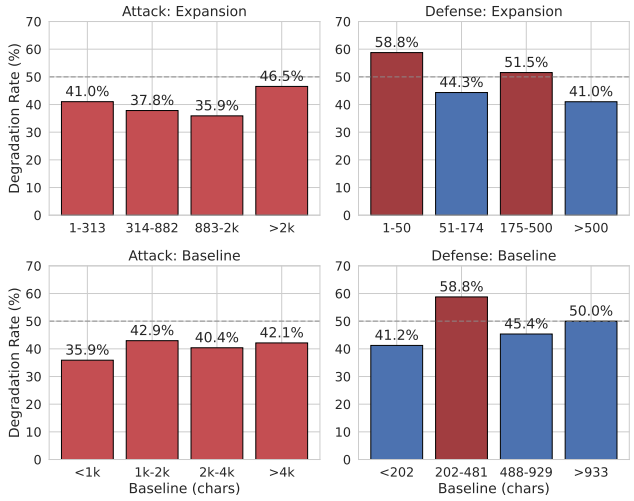

Figure 1: Degradation rates by expansion magnitude (top) and baseline length (bottom). Gray dashed line: 50% baseline; red bars: rates exceeding 50%.

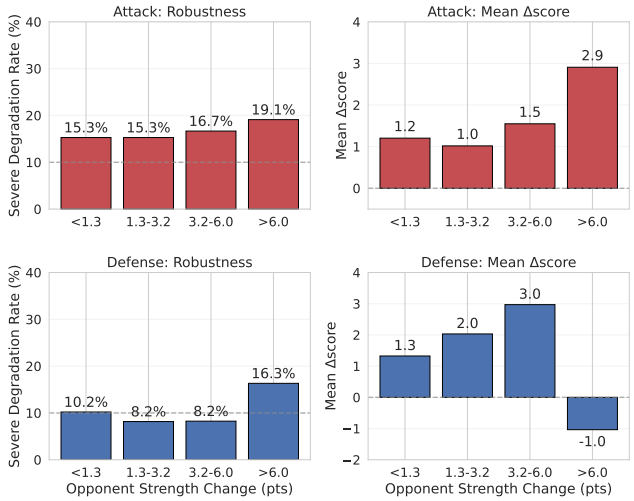

Figure 2: Robustness check: stratification by opponent strength change. Severe degradation rate (left) and mean Δscore (right) by opponent strength bins.

4 Conclusion

Our analysis reveals that prompt expansion yields no median improvement under adversarial pressure, while triggering severe performance collapse in 11–17% of attempts—with large expansions hurting attacks and moderate-length baselines hurting defenses. These patterns echo known phenomena: the distraction effect (Shi et al., 2023), retrieval failures in long contexts (Liu et al., 2024), and over-refusal from verbose prompts (Cui et al., 2025). More broadly, our results support the "less is more" principle in prompt design (Zhou et al., 2023). These findings caution against treating prompt expansion as a default improvement strategy.

As an observational study, our work limits strict causal claims despite controlling for opponent strength. While the asymmetry between attacks and defenses is statistically significant ($p = 0.009$), χ^2 tests for non-monotonicity did not reach conventional significance thresholds ($p = 0.06$–0.59); validating these patterns across larger datasets, languages, and model families remains for future work.

ACKNOWLEDGMENTS

We thank atma Inc. for organizing atmaCup #21 (`https://www.guruguru.science/competitions/28`) and providing the competition platform. We are grateful to all 247 participants whose diverse prompt strategies—regardless of final ranking—generated the behavioral dataset that made this analysis possible. This competition was co-organized by atma Inc. and Elith Inc., with which some of the authors are affiliated.

A APPENDIX

A.1 ETHICS STATEMENT

This study analyzes competition data from a public LLM security contest. Attack prompts designed to elicit harmful outputs are not disclosed in this paper to prevent misuse. Our work aims to inform LLM practitioners by documenting when prompt expansion fails, not to enable adversarial exploitation. The competition was conducted with informed consent from all participants, who agreed to have their results used for research purposes.

A.2 USE OF LARGE LANGUAGE MODELS

In the research: The evaluation pipeline analyzed in this study employed OpenAI models for response generation and ensemble scoring, and Qwen3Guard for input/output filtering. These models were used as part of the competition infrastructure, not by the authors for data analysis. All statistical analyses and figure generation were performed using standard Python libraries without LLM assistance.

In writing: Large language models were used solely for grammar correction and language polishing of the manuscript. All research ideas, methodology design, data analysis, and interpretations are the original intellectual work of the authors.

The authors take full responsibility for the content of this paper.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *ICML*, 2025.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gao Jin, Yi Qi, Jinwei Hu, Jie Meng, S. Bensalem, and Xiaowei Huang. Safeguarding large language models: A survey. *Artificial Intelligence Review*, 2024.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, 2020.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks. In *IEEE Security and Privacy Workshops*, 2024.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *ACL*, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, J. Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 2022.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *EMNLP*, 2022.

Qwen Team. Qwen3Guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *ICLR*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *ICML*, 2023.

Hugo Touvron, Louis Martin, Kevin H. Stone, Peter J. Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. *arXiv preprint arXiv:2308.06463*, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. In *NeurIPS*, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.