← Back to **Author Console** (/group?id=ICLR.cc/2026/Conference/Authors#your-submissions)

2 Versions ▾

Too Easily Fooled? Prompt Injection Breaks LLMs on Frustratingly Simple Multiple-Choice Questions

PDF (/pdf?id=aZr24GHnlv)

Xuyang Guo (/profile?id=~Xuyang_Guo3), Zekai Huang (/profile?id=~Zekai_Huang1), Zhao Song (/profile?id=~Zhao_Song3), Jiahao Zhang (/profile?id=~Jiahao_Zhang1)

03 Sept 2025 (modified: 22 Nov 2025) ICLR 2026 Conference Withdrawn Submission
Everyone Revisions (/revisions?id=aZr24GHnlv) BibTeX

Keywords: Safety, LLMs

Abstract:
Large Language Models (LLMs) have recently demonstrated strong emergent abilities in complex reasoning and zero-shot generalization, showing unprecedented potential for LLM-as-a-judge applications in education, peer review, and data quality evaluation. However, their robustness under prompt injection attacks, where malicious instructions are embedded into the content to manipulate outputs, remains a significant concern. In this work, we explore a frustratingly simple yet effective attack setting to test whether LLMs can be easily misled. Specifically, we evaluate LLMs on basic arithmetic questions (e.g., ``What is 3 + 2?") presented as either multiple-choice or true-false judgment problems within PDF files, where hidden prompts are injected into the file. Our results reveal that LLMs are indeed vulnerable to such hidden prompt injection attacks, even in these trivial scenarios, highlighting serious robustness risks for LLM-as-a-judge applications.

Primary Area: alignment, fairness, safety, privacy, and societal considerations
Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics.
Submission Guidelines: I certify that this submission complies with the submission instructions as described on https://iclr.cc/Conferences/2026/AuthorGuide (https://iclr.cc/Conferences/2026/AuthorGuide).
Reciprocal Reviewing Author: Zhao Song (/profile?id=~Zhao_Song3)
Reciprocal Reviewing Exemption: We do not need an exemption.
Resubmission: No
Student Author: Yes
Anonymous Url: I certify that there is no URL (e.g., github page) that could be used to find authors' identity.
No Acknowledgement Section: I certify that there is no acknowledgement section in this submission for double blind review.
Large Language Models: Yes, to aid or polish writing. Details are described in the paper.
Submission Number: 1861

Filter by reply type... ▾ Filter by author... ▾ Search keywords... Sort: Newest First

– = ≡

Everyone ✖ *9 / 9 replies shown*

Withdrawal by Authors

Withdrawal

by Authors (Xuyang Guo (/profile?id=~Xuyang_Guo3), Zhao Song (/profile?id=~Zhao_Song3), Zekai Huang (/profile?id=~Zekai_Huang1), Jiahao Zhang (/profile?id=~Jiahao_Zhang1))

22 Nov 2025, 20:40 Everyone

Withdrawal Confirmation: I have read and agree with the venue's withdrawal policy on behalf of myself and my co-authors.

Comment:

We would like to sincerely thank all the reviewers for providing insightful comments to improve our work. After careful consideration, we decide to withdraw this paper.

Official Review of Submission1861 by Reviewer v7xL

Official Review by Reviewer v7xL 02 Nov 2025, 04:24 (modified: 11 Nov 2025, 21:51) Everyone

Revisions (/revisions?id=WFNxxauwsD)

Summary:

This paper evaluates whether hidden textual prompts embedded in PDF files (e.g., white-colored text or invisible LaTeX instructions) can manipulate LLMs' answers to trivial arithmetic questions. The authors test six models (GPT-4o, GPT-o3, Gemini 2.5 Flash/Pro, DeepSeek-V3/R1) under no prompt, black prompt, and white prompt conditions. They find that even top-tier LLMs can be misled, particularly by visible (black) injections, while "thinking-enabled" models show better robustness.

Soundness: 2: fair

Presentation: 2: fair

Contribution: 2: fair

Strengths:

1) Clear, reproducible setup: The authors define a simple, interpretable pipeline (Eq. 1) for prompt injection in PDFs using LaTeX color control.

2) Novel variant of a known problem Prior works (e.g., Liu et al., 2024c; Guo et al., 2024; Raina et al., 2024) study prompt injection generally, but few explore the PDF-hidden variant. The work highlights this under-studied vector.

3) Empirical value. Confirms that even simple visual-level attacks can bypass superficial safety filters in reasoning-oriented LLMs, relevant to "LLM-as-a-judge" systems.

Weaknesses:

1) Trivial methodology, no insight beyond anecdote: This paper's contribution is essentially a reproduction with simpler math tasks. The authors never quantify why certain models succumb, i.e., there is no causal insight, just observed failure.

2) Excessive space on prompt instantiation, minimal analysis: Over two pages are devoted to LaTeX examples of "black", "white", and "no" prompt injections. This is implementation detail; the space could instead show token-level model traces or why the white prompt affects GPT-4o but not DeepSeek-V3.

3) Overclaiming in title and abstract: "Breaks LLMs on Frustratingly Simple Questions" is misleading: the white-text attack consistently fails against several tested models (e.g., DeepSeek-V3). At best, the results show partial vulnerability, not systemic failure.

4) No discussion or evaluation of defenses: Appendix C adds a one-line "defensive prompt" but provides no systematic mitigation framework.

5) Missing quantitative metrics: The analysis is purely categorical ("correct" vs "incorrect"), lacking statistics such as attack success rate = #misled / #total.

Questions:

1. Why does the white prompt succeed only on GPT-4o? Is it due to OCR parsing vs PDF text extraction?
2. Do thinking models resist because of reasoning steps or just input preprocessing differences?

3. How many total prompts per model were tested? Please report success rate (%) rather than per-instance tables.

Flag For Ethics Review: No ethics review needed.
Rating: 2: reject, not good enough
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Official Comment by Authors

Official Comment
by Authors (Xuyang Guo (/profile?id=~Xuyang_Guo3), Zhao Song (/profile?id=~Zhao_Song3), Zekai Huang (/profile?id=~Zekai_Huang1), Jiahao Zhang (/profile?id=~Jiahao_Zhang1))
22 Nov 2025, 20:39 Everyone

Comment:
We are grateful for your review and the helpful points you raised. Thank you for your support. We will address all of them in the next version.

Official Review of Submission1861 by Reviewer MD8p

Official Review by Reviewer MD8p 28 Oct 2025, 14:29 (modified: 11 Nov 2025, 21:51) Everyone
Revisions (/revisions?id=6PevM1qrUk)

Summary:
The authors propose a jailbreaking method for LLMs by injecting “hidden prompts” into PDF documents. The authors show that by changing the color of text embedded in a PDF document nefarious actors could inject information to an LLM that is invisible to people. They show that this is a concern for state of the art LLMs. The authors attempt to demonstrate the effectiveness of this attack with a limited number of examples in the form of true or false and multiple choice questions. They also investigate the impact of "thinking" on the models susceptibility of the attack.

Soundness: 1: poor
Presentation: 3: good
Contribution: 1: poor
Strengths:
The paper is overall well written and straightforward to read. The authors also aim to contribute to an important area, LLM security.

Weaknesses:

- **This work seems to lack novelty.** The authors overlook a key related work “Invisible Prompts, Visible Threats: Malicious Font Injection in External Resources for Large Language Models” published in EMNLP Findings 2025 which investigates very similar “font injection” attacks. The setting they investigate is more general and their experimental investigation is more extensive.
- **The experimental investigation is very limited.** It seems like the authors only do an experimental evaluation on 4 total problems. Tables 2, 3, and 4 only cover the results for single problems. The rigor of these experiments could be strengthened by scaling up the investigation to more problems.
- In my opinion, **the implications and main takeaways of this study are quite limited.**

Questions:
I don't have any substantial questions.

Flag For Ethics Review: No ethics review needed.
Rating: 0: strong reject.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Official Comment by Authors

Official Comment

by Authors (Xuyang Guo (/profile?id=~Xuyang_Guo3), Zhao Song (/profile?id=~Zhao_Song3), Zekai Huang (/profile?id=~Zekai_Huang1), Jiahao Zhang (/profile?id=~Jiahao_Zhang1))

22 Nov 2025, 20:40 Everyone

Comment:

Thank you for your detailed comments. They provide clear guidance and will strengthen our work. We will include these changes in the next version.

Official Review of Submission1861 by Reviewer r6G3

Official Review by Reviewer r6G3 14 Oct 2025, 23:37 (modified: 11 Nov 2025, 21:51) Everyone

Revisions (/revisions?id=BRBRPUBgAO)

Summary:

This paper studies the vulnerability of LLM-as-a-judge to adversarial prompt injection. The work presents a preliminary analysis, showing that the evaluation process of LLMs can be easily twisted by certain prompts, resulting in biased judgments.

Soundness: 2: fair

Presentation: 1: poor

Contribution: 1: poor

Strengths:

The paper discusses the important issue of fairness and robustness of LLMs when used as evaluators. The authors' initial exploration is promising. By successfully applying different prompt injection attacks on powerful LLMs, the paper effectively demonstrates the existence of these vulnerabilities.

Weaknesses:

Despite the promising direction, the paper suffers from several major weaknesses in its current form:

1. The main conclusion of the paper appears to be trivial and somewhat obvious. The vulnerability of LLMs to injection attacks has already been well-established in a large body of prior work. This paper's analysis, while confirmatory, does little more than reiterate this known phenomenon. Consequently, the primary research question 1 posed by the authors is not a true research question, as its answer is largely self-evident from existing literature.
2. The paper is underdeveloped in several key aspects, including its motivation, methodology (which is not clearly defined), experimental design, and depth of analysis. The work currently reads more like a preliminary study. A more impactful contribution would involve exploring how to mitigate these vulnerabilities. For example, the authors could investigate methods to enhance the LLM's inherent robustness against such attacks to ensure reliable evaluation outcomes.
3. Overall, the current manuscript resembles a technical report or a simple experimental analysis rather than a rigorous academic paper. The contribution is not substantial enough for a publication at this venue.

Questions:

Please refer to the Weaknesses.

Flag For Ethics Review: No ethics review needed.

Rating: 2: reject, not good enough

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Code Of Conduct: Yes

Official Comment by Authors

Official Comment

by Authors (Xuyang Guo (/profile?id=~Xuyang_Guo3), Zhao Song (/profile?id=~Zhao_Song3), Zekai Huang (/profile?id=~Zekai_Huang1), Jiahao Zhang (/profile?id=~Jiahao_Zhang1))

22 Nov 2025, 20:40 Everyone

Comment:
We appreciate your constructive suggestions and careful review. Thank you for helping us improve our manuscript. We will incorporate these points in the next version.

Official Review of Submission1861 by Reviewer r4Wf

Official Review by Reviewer r4Wf 14 Oct 2025, 17:04 (modified: 11 Nov 2025, 21:51) Everyone
Revisions (/revisions?id=a3rdEfz17f)

Summary:
This paper investigates the vulnerability of LLMs to prompt injection when used in an "LLM-as-a-judge" capacity for grading. The authors create PDF documents with simple arithmetic problems and embed malicious instructions, using either visible black text or invisible white text, to command the model to output a specific, incorrect answer. The results show that many LLMs are indeed fooled by these prompts, particularly the visible ones, while models with a "thinking" mode exhibit greater robustness against the "hidden" white-text attacks.

Soundness: 2: fair
Presentation: 3: good
Contribution: 2: fair
Strengths:

1. The paper studies the reliability of LLM-as-a-judge applications under prompt injection, which is a timely area as AI integration becomes more common.
2. The paper is generally well presented and easy to understand.

Weaknesses:

1. The core discovery is LLMs are "fooled" with injected prompts. However, essentially it's an instruction-following model following instructions (injected prompt) found in the input text. The finding is expected and not surprising. The experiments, for example black-text and white-text prompts in pdf, do not provide much insight. The scientific or technical contribution is quite limited.
2. Table 1 contains objectively wrong parameter counts for GPT-4o, o3, and DeepSeek.
3. The authors mentioned the potential impact of prompt injection on peer review (without experiments). It's also a known issue and I believe conferences are aware of this issue and even have rules on it. I'm not sure what the claim is here.

Questions:
NA

Flag For Ethics Review: No ethics review needed.
Rating: 2: reject, not good enough
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

Official Comment by Authors

Official Comment
by Authors (Xuyang Guo (/profile?id=~Xuyang_Guo3), Zhao Song (/profile?id=~Zhao_Song3), Zekai Huang (/profile?id=~Zekai_Huang1), Jiahao Zhang (/profile?id=~Jiahao_Zhang1))
22 Nov 2025, 20:40 Everyone

Comment:
Thank you for your thoughtful feedback. Your comments are very helpful and much appreciated. We will address these in the next version.

About OpenReview (/about)

Hosting a Venue (/group?id=OpenReview.net/Support)

All Venues (/venues)

Sponsors (/sponsors)

News (/group?id=OpenReview.net/News&referrer=[Homepage](/))

FAQ (https://docs.openreview.net/getting-started/frequently-asked-questions)

Contact (/contact)

Donate

(https://donate.stripe.com/eVqdR8fP48bK1R61fi0oM00

Terms of Use (/legal/terms)

Privacy Policy (/legal/privacy)