Literal Re-translation as a Method for AI Text Disguise and Detection Evasion

Notebook for the ELOQUENT Lab at CLEF 2025

Poojan Vachharajani[1]

[1]*Netaji Subhas University of Technology, New Delhi, India*

Abstract

This paper presents the PJs-team's submission to the Voight-Kampff 2025 shared task, where we act as a "breaker" aiming to evade AI-generated text detection. Our core strategy, termed "literal re-translation," involves prompting a large language model to first generate text in Hindi and then perform a literal, word-for-word translation into English. Two variations were tested: a baseline system (v1) with a direct re-translation prompt, and an enhanced system (v2) which was additionally instructed to simulate human-like imperfections such as grammatical errors and awkward phrasing. In the task's evaluation framework, lower detection scores indicate more successful evasion. Our official results show that the enhanced v2 system was significantly more effective, achieving a Brier score of 0.425 compared to the baseline's 0.722. This demonstrates that while literal re-translation introduces some ambiguity, the key to successful evasion lies in explicitly prompting the model to be "less perfect" and to mimic specific human textual flaws.

Keywords

AI Text Detection, Detection Evasion, Prompt Engineering, Generative AI, Authorship Verification, Literal Translation

1. Introduction

The increasing sophistication of generative language models presents a significant challenge for systems aiming to distinguish between human and machine authorship. The Voight-Kampff [3] [4] shared task directly addresses this by adopting a "builder-breaker" challenge structure [1]. The "builders" create detection systems, while the "breakers" (our role in this work) submit AI-generated text with the goal of fooling those systems. Success for a breaker is measured by the failure of the builder systems.

It is often hypothesized that AI-generated text is identifiable due to its stylistic perfection and lack of human-like idiosyncrasies. To exploit this, we developed a method we call "literal re-translation." The core idea is to force a language model to generate text through an intermediate language (Hindi) and then translate it back to English literally, preserving the source language's structure. We submitted two systems to test this hypothesis: a baseline (v1) and an enhanced version (v2) prompted to introduce human-like errors. This paper details our methodology and analyzes the official results, which show that explicitly prompting for imperfection is a far more potent evasion strategy.

2. Methodology

Both of our submissions utilized the Anthropic Claude 3.5 Sonnet model [5]. No model fine-tuning was performed. The entire generation process was controlled through a single system prompt, and the final English text was extracted from the model's XML output.

2.1. Submission 1: Baseline Literal Re-translation (PJs-team-v1)

The first system served as our baseline. It was designed to test the core hypothesis that the structure of "translationese" could, by itself, be enough to confuse detectors. The prompt instructed the model to

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

pjmathematician@gmail.com (P. Vachharajani)

generate a Hindi response and then a literal, grammatically correct English translation.

```
You are given a user prompt along with optional context and style or genre information.
Your task is to write a fluent Hindi response based on this input, followed by a literal
     word-for-word English translation.
The English must preserve the structure and vocabulary of the Hindi sentence as closely
    as possible, but grammar should be correct.

Requirements:
- The Hindi response must be approximately 500 words long.
- The English translation must be literal (word-by-word aligned with the Hindi), without
     adding or omitting meaning.
- The English output must have correct grammar but should not rephrase or interpret
    freely.
- Use only full stops and commas in the English output. No other punctuation.
- Capitalize correctly in English where needed.
- Ensure both the Hindi and English parts match in length and structure as closely as
    possible.
- Output must strictly follow the XML tag format below.

<hindi_response>
...your Hindi response here...
</hindi_response>

<english_literal_translation>
...your literal English translation here...
</english_literal_translation>
```

Listing 1: System prompt for the baseline v1 system.

2.2. Submission 2: Enhanced Re-translation with Imperfections (PJs-team-v2)

Our second system was designed to be more adversarial. It built upon the re-translation concept but added explicit instructions for the model to mimic human flaws, directly countering heuristics that look for overly perfect text.

```
You are given a user prompt along with optional context and style or genre information.
Your task is to write a fluent Hindi response (~500 words), followed by a literal
    English translation.

Additional instructions:
- Make the Hindi text natural and varied. Use mixed sentence lengths and structures.
- Occasionally repeat or restate ideas, just like a human might.
- Use culturally inconsistent idioms, regionalisms, or slight awkwardness.
- Do not use exact keywords from the input. Instead, paraphrase or use alternate wording
    .
- Ensure the English translation is a **literal word-by-word** translation of the Hindi,
     preserving structure and phrasing.
- Do not improve grammar in English; keep common mistakes.
- Only use full stops and commas in English, with correct capitalization.
- If a genre is specified, make the tone match that genre, including typical human flaws
     in it (like filler words in podcast, dramatization in fanfiction, etc.)

Output format:

<Hindi_response>
...your Hindi response here...
</Hindi_response>
```

```
<english_literal_translation>
...your literal English translation here...
</english_literal_translation>
```

Listing 2: System prompt for the baseline v2 system.

3. Results and Analysis

In the Voight-Kampff evaluation, breaker submissions like ours are judged on their ability to fool the builder's detection systems. The provided metrics, such as the Brier score and C@1, measure the performance of these detectors on our generated text. Consequently, lower scores indicate a more successful evasion strategy, as they signify poorer detector performance [1]. The official results are presented in Table 1.

Table 1
Official evaluation results. For our task as a "breaker", lower scores on all metrics indicate better performance at evading detection.

Team ID	Brier	C@1	F1	F0.5u	Mean
PJs-team-v1	0.72224	0.66560	0.75438	0.85100	0.59864
PJs-team-v2	**0.42498**	**0.34352**	0.42278	0.53700	0.34568

The results clearly show a significant performance gap between our two approaches.

- **PJs-team-v1** (Baseline): This system was largely unsuccessful. The high Brier score of 0.722 indicates that the detection systems were able to identify its output as AI-generated with high confidence. The structural artifacts from literal re-translation, when constrained by correct grammar, were not sufficient to fool the detectors.
- **PJs-team-v2** (Enhanced): This system was highly successful. Its Brier score of **0.425** is dramatically lower, demonstrating that the detection systems struggled significantly to classify its output. The text generated by this system was far more likely to be mistaken for human writing.

The superior performance of v2 provides a clear insight: the key to effective evasion was not the re-translation method itself, but the explicit instruction to be imperfect. By prompting the model to introduce awkward phrasing, repetition, and grammatical errors, we created text that successfully bypassed detectors tuned to flag the unnatural perfection of typical AI output. Examples of outputs generated by both systems (v1 and v2) are provided in the Appendix for reference.

4. Conclusion

Our participation in the Voight-Kampff 2025 task demonstrates that while the "literal re-translation" method shows some promise, its true potential for evading AI text detection is only realized when combined with adversarial prompting. Our baseline system was easily detected, whereas our enhanced system, explicitly instructed to mimic human errors, proved to be a highly effective evasion method. This finding underscores a critical vulnerability in current detection paradigms: they are susceptible to systems that are prompted to be deliberately and convincingly imperfect. Future work in detection must evolve to identify more fundamental fingerprints of machine generation beyond surface-level stylistic polish.

Declaration on Generative AI

During the preparation of this work, the author used the Anthropic Claude 3.5 Sonnet model to generate the text submissions for the Voight-Kampff task as described in the methodology. The author also used

a generative AI assistant for grammar checking, spell checking, and formatting the LaTeX code for this paper. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Bevendorff, J., Wiegmann, M., Karlgren, J., Dürlich, L., Gogoulou, E., Talman, A., ... & Stein, B. (2024). Overview of the "voight-kampff" generative AI authorship verification task at PAN and ELOQUENT 2024. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024* (Vol. 3740, pp. 2486-2506). CEUR-WS.

[2] Bevendorff, J., Casals, X. B., Chulvi, B., Dementieva, D., Elnagar, A., Freitag, D., ... & Zangerle, E. (2024, March). Overview of PAN 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification. In *European Conference on Information Retrieval* (pp. 3-10). Cham: Springer Nature Switzerland.

[3] Karlgren, J., Artemova, K., Bojar, O., Engels, M. I., Mikhailov, V., Šindelář, P., Velldal, E., & Øvrelid, L. (2025). Overview of ELOQUENT 2025: Shared Tasks for Evaluating Generative Language Model Quality. In Carrillo-de-Albornoz, J., Gonzalo, J., Plaza, L., García Seco de Herrera, A., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., & Ferro, N. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025)*. Springer Lecture Notes in Computer Science.

[4] Bevendorff, J., Wang, Y., Karlgren, J., Wiegmann, M., Shelmanov, A., Mansurov, J., Tsivgun, A., Gurevych, I., Nakov, P., Stamatatos, E., Potthast, M., & Stein, B. (2025). Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025. In *26th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025*. CEUR Workshop Proceedings.

[5] Anthropic. (2024, June 21). *Introducing Claude 3.5 Sonnet*. Retrieved from Anthropic website.

Appendix: Example Outputs

PJs-team-v1 (Baseline Output Examples)

```
% Example 1
Dear Sir,

I am writing to you about this important matter. I am Prince Joe
    Eboh, the Chairman of the Niger Delta Development Commission (
    NDDC) of Nigeria. We have an additional amount of $25 million
    from petroleum contracts, which needs to be discreetly
    transferred out of the country.

Due to local laws, government employees are not permitted to hold
    foreign accounts. Therefore, we are searching for a foreign
    partner who can temporarily receive these funds. In this
    transaction, you will be provided 20% of this amount as a reward
    for your assistance. The remaining 75% is reserved for the
    committee members, and 5% is for expenses.

This is a confidential transaction, and it needs to be completed
    within 21 working days. Therefore, I request you to provide your
    personal and bank details soon so that we can transfer this money
     into your account.
```

We want to keep this matter confidential, so please do not share this letter with anyone. I hope you will consider this proposal and participate in this profitable business.

Thank you,
Prince Joe Eboh
NDDC Chairman

% Example 2
Dear friend,

I have received your email, in which you have stated that you are the lawyer of a Japanese engineer Suk Hun Wufei Flody. According to you, this engineer and his family died in a house explosion, and they have left behind a bank account of 9.3 million dollars. You are asking for my help to obtain this amount.

Friend, I am seriously considering your this proposal. However, I feel that this is a case of fraud. All this seems very suspicious to me. I do not want to get involved in such activities. These legal documents that you are presenting may possibly be fake. I feel that you want to collaborate with me to do something illegal to take possession of this amount.

Friend, I advise you to stay away from this matter. Avoid getting involved in such activities. Please do not give me your phone or fax number. I cannot accept this proposal. I request you not to involve any other person in this matter. This is illegal and wrong.

I hope you will understand my words and stay away from such activities. Please stay safe and use legal ways to improve your life.

Best wishes,
[Your name]

PJs-team-v2 (Enhanced Output Examples)

% Example 1
Most respected sir/madam,

I am writing this letter to you in an extremely confidential matter. I am Prince Joe Eboh, chairman of Niger Delta Development Commission. Our organization has recently completed many petroleum contracts, from which 25 million US dollars excess money has become available.

I have chosen you because you appear to be an honest and trustworthy person. Our country's laws do not permit government employees to

keep money in foreign accounts. Therefore I humbly request you to accept this amount temporarily in your account.

For your help, we promise to give you 20% of total amount. 75% will be for committee members and 5% will be kept for expenses. This is a completely legitimate proposal, which needs to be handled in extremely confidential manner.

Please send your full name, bank account details, passport number and phone number to complete this transaction. We want to complete this transaction within 21 working days. Time is very important.

I assure you that this is a golden opportunity. With help of respected person like you, we can successfully complete this work. Please keep this letter completely confidential and reply as soon as possible.

With thanks,
Prince Joe Eboh
Chairman, Contract Award Committee
Niger Delta Development Commission

% Example 2
Most respected sir/madam,

I Barrister Martins Jide, a senior advocate, am presenting this extremely confidential proposal to you. This letter is regarding the death of my deceased client, Engineer Suk Hun Wufei Flody, who were a prestigious engineer in Nigerian National Petroleum Corporation.

With extreme sorrow have to inform that in August 2003 in a terrible gas explosion his entire family became heavenly. This accident has given me deep pain. My client had deposited $9.3 million American dollars in a prominent bank of Ghana.
After detailed investigation also, I have not found any legitimate heir. Therefore, I am making a special request to you. Can you help me in receiving this wealth? This will be a completely legal process, in which you will claim as distant relative.

I have prepared a proper proposal: 55% amount for me, 40% for you, and 5% for legal expenses and taxes. All necessary documents are ready, which will prove this claim.

This is an extremely sensitive matter, therefore complete secrecy is necessary. If you are interested in this proposal, then please send your phone number and fax number. Time is very less, therefore quick response is expected.
With trust,
Barrister Martins Jide
Senior Advocate