Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29099

Kaomoji Fixed Translation as Knowledge Hallucination in LLMs: A Case Study on XiaoHongShu

Jinhan Feng

Nanjing Jinling High School, Nanjing, China
fengjinhan1013@gmail.com

Abstract. Large Language Models (LLMs) produce fluent but sometimes unfounded outputs, a phenomenon commonly called hallucination. On the XiaoHongShu (RED) platform, when users input kaomoji—ASCII or Unicode emoticons—the translation tool often returns a stable, seemingly meaningful Chinese phrase even though the input lacks explicit semantic content. This paper examines why LLMs generate such fixed translations. Building on the concept of hallucination snowballing, classifications of hallucination types, and methods for reducing knowledge hallucinations, through case analysis, literature synthesis, and mechanistic review, this paper mainly discusses: (1) LLMs produce consistent translations for semantically null inputs. (2) Which hallucination category best fits this case? (3) How do prompt framing, pretraining co-occurrence patterns, and autoregressive decoding contribute? This study argues that the kaomoji fixed translation is primarily a form of knowledge hallucination reinforced by prefix consistency and statistical co-occurrence. This paper concludes by recommending uncertainty-aware behaviors, prompt-level checks, and data interventions to reduce such errors.

Keywords: Large Language Models, knowledge hallucination, kaomoji, fixed translation, consistency bias

1. Introduction

Large Language Models (LLMs) such as GPT, PaLM, and LLaMA have greatly improved machine translation and conversational agents, but their outputs can be untrustworthy when they hallucinate —that is, when they state fluent yet unsupported claims [1]. Hallucinations arise across tasks, including summarization, open-domain QA, and dialogue, prompting researchers to categorize and study different types of errors [2-4]. Zhang et al. described a “hallucination snowball” in which an early incorrect assertion causes subsequent outputs to remain consistent with that incorrect assertion [5]. Lü et al. focused on knowledge hallucinations and methods to reduce them through retrieval and context highlighting [1]. Wu et al. considered relation and visual hallucinations, showing that models sometimes favor commonsense priors over the actual input [6]. This paper investigates a text-only phenomenon observed on XiaoHongShu: when users enter kaomoji as the entire input to the platform’s translation tool, the model returns apparently fixed Chinese translations. For example, (ﾒ￣ω￣ﾒ) is often rendered as “concerning” and (•̀へ•́╬)” as “angry,” with little variation across trials. These outputs are not grounded in a shared linguistic meaning of the symbols; instead, they

DOI: 10.54254/2755-2721/2026.TJ29099

reflect the model's learned associations. The goals are to: (1) classify the phenomenon within hallucination taxonomies; (2) analyze the mechanisms—prompting, pretraining associations, and decoding—behind the behavior; and (3) propose practical mitigations. To accomplish these goals, this paper combines hands-on case analysis, a review of relevant literature on text-based hallucinations, and a mechanistic walkthrough of generation behavior.

2. Hallucination taxonomy in text-based LLMs

Hallucinations in language models are often separated into categories that help diagnosis and mitigation. A commonly used category is knowledge hallucination, where the model asserts factual claims that are not supported by evidence [2]. Relation hallucination occurs when a model fabricates or misattributes links between entities—a phenomenon often examined in multimodal settings but equally relevant to text-only scenarios where associations lack justification [6]. The hallucination snowball effect describes how an initial error is preserved and amplified as the model strives for coherence with its prior output [5]. Moreover, large language models display a consistency bias and lexical entrenchment: under identical prompts and decoding parameters—especially with deterministic decoding—they repeatedly produce the same or nearly identical responses [7,8].

3. The kaomoji fixed translation phenomenon

This paper performed controlled queries on the XiaoHongShu translation interface and on other web-based translator instances that use similar LLM backends. In the XiaoHongShu tests, we repeatedly submitted single kaomoji characters or sequences and recorded the outputs. Across 30 independent trials for several kaomoji, the system returned virtually the same target phrase in Chinese. Examples collected include: (灬ꈍ ꈍ灬) → "concerning~"; (•̀へ•́╬) → "angry"; ¯_(ツ)_/¯ → "don't care". Despite their lack of fixed lexical meaning, the kaomoji inputs yield output strings that are immediately readable as affection, anger, or indifference. This study also compared different prompt wrappers—explicitly asking for a translation versus asking for an interpretation—and found the translation framing increased the likelihood of a fixed, single-phrase output [9].

4. Mechanisms behind the phenomenon

Several interacting mechanisms explain the emergence of fixed translations for kaomoji.

(1) Pretraining Co-occurrence and Statistical Association. LLMs learn token associations from massive, noisy web corpora that include social media posts, chat logs, and forum discussions. Kaomoji and other emotive symbols frequently co-occur with short affective descriptions or reaction words in such corpora. Studies of emoji and emoticon semantics show that pictographic and typographic emoticons have strong statistical ties to sentiment-bearing words [10]. As a result, when a model sees a kaomoji in isolation and is asked to translate, it relies on learned co-occurrence patterns to pick the most probable corresponding phrase.

(2) Prompt Framing and Task Constraint. Translation prompts create a task demand: the model must render the input into a target language. Even when the input is non-linguistic, the model attempts to comply. The implicit prompt typically resembles "translate the following into Chinese: [input]". Under this instruction-following pressure, the model prefers producing a concise, high-probability mapping from input to output rather than returning a refusal or a request for clarification [9].

DOI: 10.54254/2755-2721/2026.TJ29099

(3) Autoregressive Decoding and Prefix Commitment. Modern LLMs generate tokens sequentially; once the model outputs an initial token of the translation, subsequent tokens become conditioned on that prefix. This "prefix commitment" effect reduces the chance of backtracking to an alternative interpretation and encourages consistent completions [11]. Deterministic or low-temperature decoding further increases repeatability, making identical inputs yield identical outputs [7].

(4) Consistency Bias and Caching. If a model is used in production, logging, caching, or fine-tuning on user interactions may entrench early mappings. An initial mapping that is frequently returned and logged may become reinforced via administrative fine-tuning or retrieval caches. This creates a feedback loop resembling the hallucination snowball described by Zhang et al. [5].

(5) Absence of Uncertainty Representation. Most off-the-shelf LLMs are not trained to output calibrated uncertainty measures when a token is non-linguistic. Without a mechanism to signal "no translation", the model defaults to its most likely verbalization based on its training distribution [12].

5. Related work on text-based hallucinations

Textual hallucination has been widely studied in summarization and QA. Maynez et al. documented unsupported assertions in abstractive summaries and argued that model training objectives can promote fluent but unfaithful generations [2]. Pagnoni et al. investigated factuality metrics and methods for measuring unsupported content [13]. Shuster et al. explored retrieval-enhanced generation as a means to ground outputs, while Ji et al. surveyed broader hallucination taxonomy and mitigation techniques [3,4]. Other work has focused on prompt sensitivity and how different instruction patterns change hallucination rates [9]. Countermeasures such as retrieval augmentation, chain-of-verification, and highlighting of key reference passages (e.g., COFT) have shown promise, but they assume that the input has verifiable external facts to retrieve or highlight [1]. When the input is non-linguistic, these approaches are less directly applicable, because there may be no external evidence to fetch or highlight.

6. Mitigation strategies

This study identifies several practical mitigation strategies tailored to the kaomoji fixed-translation problem. (1) Explicit Non-translatable Detection. Add a lightweight classifier that first determines whether the input is linguistic and translatable. If the classifier flags the input as non-linguistic, the system should avoid forced translation and instead return a neutral response (e.g., "No direct translation available") or ask for clarification [14]. (2) Confidence and Refusal Mechanisms. Train or prompt the LLM to produce a calibrated uncertainty estimate or a refusal when confidence is low. Prompt-based techniques can ask the model to self-evaluate the translation probability and avoid output if below a threshold [15]. (3) Data-level Corrections. During dataset curation, reduce strong co-occurrence signals that map kaomoji directly to short words by annotating or filtering social-media-derived pairs. Data attribution and filtering techniques can help mitigate spurious association learning [16]. (4) Controlled Decoding. Use higher temperature or stochastic sampling, or enforce diverse beam outputs combined with a verification stage, to avoid a single entrenched mapping. However, this reduces determinism and may not be suitable for production unless combined with verification [7]. (5) Logging and Human-in-the-Loop Review. If rare or surprising mappings are detected in logs, flag them for human review and corrective action in the training pipeline. This breaks the cache-and-fine-tune reinforcement loop that can entrench a bad mapping [5].

DOI: 10.54254/2755-2721/2026.TJ29099

7. Conclusion

This paper analyzed a case where LLM-based translation on XiaoHongShu consistently maps kaomoji to fixed Chinese phrases. We classified this phenomenon primarily as a knowledge hallucination reinforced by prefix consistency and statistical co-occurrence learned during pretraining [5]. The behavior is exacerbated by translation prompts that force the model to provide an output even for non-linguistic inputs, and by deterministic decoding that locks in the first plausible completion [9,11]. Logging, caching, and production fine-tuning can further amplify the mapping, producing a snowball-like effect [2].

Addressing this class of hallucination requires systems that can detect non-linguistic inputs and adopt uncertainty-aware or refusal behaviors. Data curation and annotation strategies can reduce spurious co-occurrence learning, while prompt-level and decoding interventions can decrease the probability of a single, entrenched mapping. Beyond immediate engineering solutions, this case highlights a broader research gap: most hallucination mitigation techniques assume the presence of factual content to verify, but non-semantic inputs like kaomoji expose a blind spot in current LLM training and evaluation.

Future research should therefore explore calibrated refusal signals, better pretraining curation for social-media-like content, and lightweight classifiers to detect non-translatable inputs. Building such mechanisms will not only improve translation reliability but also enhance overall user trust in AI systems. In this sense, the kaomoji phenomenon serves as a concrete reminder that hallucination is not just a factuality problem—it is also a problem of context framing, task design, and user interaction.

References

[1] Lü, Q., et al. (2024) Coarse-to-Fine Highlighting: Reducing Knowledge Hallucination in Large Language Models. Proceedings of the International Conference on Machine Learning (ICML).
[2] Maynez, J., et al. (2020) On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).
[3] Shuster, K., et al. (2021) Retrieval-Enhanced Generative Models. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).
[4] Ji, Z., et al. (2023) Survey of Hallucination in Natural Language Generation. ACM Computing Surveys.
[5] Zhang, M., et al. (2024) How Language Model Hallucinations Can Snowball. Proceedings of the 41st International Conference on Machine Learning (ICML).
[6] Wu, M., et al. (2024) Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models. Proceedings of the International Conference on Machine Learning (ICML).
[7] Holtzman, A., et al. (2020) The Curious Case of Neural Text Degeneration. Proceedings of the International Conference on Learning Representations (ICLR).
[8] Pickering, M., & Garrod, S. (2020) Toward a Mechanistic Psychology of Dialogue. Behavioral and Brain Sciences.
[9] Lee, N., et al. (2022) Prompt Sensitivity in Large Language Models. arXiv: 2212.10559.
[10] Barbieri, F., et al. (2018) Modelling the Semantics of Emoji. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
[11] Vaswani, A., et al. (2017) Attention is All You Need. Proceedings of the Conference on Neural Information Processing Systems (NeurIPS).
[12] Evans, O., et al. (2021) TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).
[13] Pagnoni, A., et al. (2021) Understanding Factuality in Abstractive Summarization. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).
[14] Lin, S., et al. (2022) Teaching Models to Refuse Unknowns. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

DOI: 10.54254/2755-2721/2026.TJ29099

[15] Zhao, Z., et al. (2023) Revisiting Chain-of-Thought Reasoning. Proceedings of the Conference on Neural Information Processing Systems (NeurIPS).
[16] Kim, B., et al. (2023) Reducing Hallucination via Data Attribution. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).