HalloMTBench: A Large-Scale Human-Verified Benchmark for Diagnosing Hallucination Triggers in LLM Translation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have advanced Machine Translation (MT), yet they remain vulnerable by hallucinations. Most existing benchmarks, originally designed for neural MT models, do not effectively elicit hallucinations from multilingual LLMs, resulting in a low density of samples for LLM translation hallucination analysis. To address this, we create HalloMTBench, a human-verified hallucination benchmark derived from tens of millions of translations, for studying signals that trigger LLM translation hallucinations across 11 English-to-X directions. We curate 6,908 authentic hallucination instances through a rigorous process: frontier LLMs first translate a large source corpus, and their outputs are then validated by both an LLM-judge ensemble and human experts. Based on our analysis of these hallucination instances, we introduce a taxonomy for translation hallucination that distinguishes **Instruction Detachment** from **Source Detachment**. Our evaluation of 21 model variants on HalloMTBench reveals several observable vulnerability patterns, including model Characteristics, source length, language proximity, hallucination reproducibility, and model-dependent effects of reasoning-style decoding. HalloMTBench offers a forward-looking testbed for diagnosing LLM translation hallucination. All resources will be released.

1 Introduction

Large Language Models have demonstrated remarkable progress in machine translation (Zhao et al., 2023; Chang et al., 2024), often surpassing traditional neural machine translation (NMT) systems in fluency and contextual understanding (Zhang et al., 2023; Zhu et al., 2023, 2024). However, their practical deployment remains significantly hampered by translation hallucination, where models generate output that is nonsensical or unfaithful to source text (Guerreiro et al., 2023a; Huang et al., 2025; Gogoulou et al., 2025).

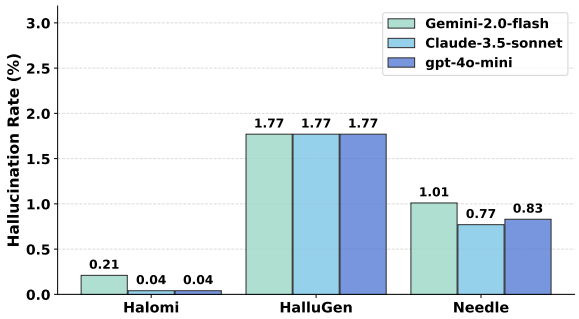


Figure 1: Existing machine translation hallucination benchmarks yield too few hallucination cases for current LLMs. Frontier LLMs demonstrate near-zero hallucination rates across three prominent benchmarks.

Addressing LLM translation hallucinations faces two critical obstacles. First, while prior work has categorized hallucinations in NMT (Guerreiro et al., 2023a; Zhang et al., 2025), the autoregressive and instruction-following nature of LLMs introduces novel and complex hallucination types. This results in the lack of a definitional framework specifically tailored for LLM-induced translation hallucinations. Second, existing hallucination evaluation benchmarks. As demonstrate in our preliminary analysis (Figure 1), frontier LLMs, including `Claude-3.5`, `Gemini-2.0`, and `GPT-4o`, exhibit near-zero hallucination rates (under 2%) on prominent MT hallucination benchmarks like Halomi (Dale et al., 2023), HalluGen (Dürlich et al., 2024), and Needle (Guerreiro et al., 2023b). This sparsity of discoverable hallucination instances severely impedes in-depth analysis of translation hallucinations.

To bridge this gap, we first propose a new taxonomy for LLM translation hallucinations. Referencing the definition of faithfulness-based hallucinations (Huang et al., 2025), we categorize translation hallucination into two primary classes: **In-**

struction Detachment, which encompasses hallucination to adhere to task instructions (e.g., producing untranslated content or generating output in the wrong target language), and **Source Detachment**, referring to deviations from the source content (e.g., fabricating extraneous information or generating repetitive text). This taxonomy provides a clear and actionable lens for judging LLM translation behaviors.

Guided by this taxonomy, we curate **HalloMT-Bench**, a large-scale, human-verified benchmark spanning 11 English-to-X translation directions. The benchmark is curated through a rigorous four-stage pipeline designed for both scale and reliability (see Figure 3). We began by generating over 16M translations from a 4M-sentence corpus using several frontier LLMs. Next, an LLM-judge ensemble performs automated hallucination detection. These are then submitted for expert annotation by professional linguists. Finally, a quality control loop with senior annotators resolve ambiguities, yielding the final set of **6,908** instances. We conduct an extensive evaluation on HalloMTBench across 21 frontier models, including 14 closed- and 7 open-source LLMs. Our results establish an initial leadboarder for hallucination based on this dataset, revealing a wide performance spectrum, with `HY-MT1.5-7B` exhibiting the lowest rate and `Hunyuan-MT-7B` the highest.

Further analysis reveals potential **hallucination triggers**, which we treat as correlational patterns rather than direct causes. Key findings include: Source Length: hallucination rates follow a U-shaped pattern with source length. Language Proximity: linguistically or orthographically related languages are more prone to language-mixing errors. Reproducibility and Reasoning Modes: many hallucinations persist under re-generation, while thinking-style decoding increases token usage and yields model-dependent hallucination changes. We further provide case studies in Appendix A.6, and will release all resources.

In summary, our contributions are threefold:

- We construct and release **HalloMTBench**, a large-scale, multilingual, and human-verified benchmark designed to elicit and evaluate hallucinations in modern LLMs.
- We propose a fine-grained taxonomy for analyzing LLM translation hallucination, which distinguishes between Source Detachment and Instruction Detachment.
- Through extensive evaluation, we identify potential hallucination triggers and robustness patterns in LLM translation.

2 Related Work

Hallucination in Machine Translation The study of hallucinations in the Neural Machine Translation (NMT) era laid foundational groundwork (Xiao et al., 2023; Tang et al., 2025), establishing a taxonomy that broadly categorized them into two main types: hallucinations under perturbation and natural hallucinations (Guerreiro et al., 2023a; Raunak et al., 2021). Hallucinations under perturbation arise when minor, artificial disturbances to the source text (e.g., spelling errors) provoke severely degraded or nonsensical output (Lee et al., 2018). In contrast, natural hallucinations, which occur with clean, unperturbed inputs, are harder to detect. These are further subdivided into *oscillatory hallucinations*, characterized by repetitive and meaningless word or phrase generation, and *detached hallucinations*, where the output is fluent yet semantically disconnected from the source. A subsequent, more fine-grained analysis delineated detached hallucinations into *strongly detached* (partially related to the source) and *fully detached* (completely unrelated) variants (Guerreiro et al., 2023b).

Hallucination in LLMs The advent of LLMs has broadened the scope of hallucination to encompass a wider array of hallucination modes, reflecting their distinct architectures and training paradigms (Ji et al., 2023; Tonmoy et al., 2024; Huang et al., 2025). A prominent distinction is drawn between intrinsic and extrinsic hallucinations (Gogoulou et al., 2025; Zhou et al., 2021): intrinsic hallucinations are attributed to flawed reasoning capabilities of the model, rather than knowledge deficits, while extrinsic hallucinations stem from factual inaccuracies or gaps in parametric knowledge. A related, yet distinct dichotomy concerns factuality versus faithfulness (Gogoulou et al., 2025). Factuality-based hallucinations involve discrepancies between generated content and verifiable real-world facts, whereas faithfulness-based hallucinations describe divergences from the user's input or a lack of internal consistency within the output.

Hallucination Benchmarks for Translation Specialized MT hallucination benchmarks have

Table 1: Summary of major existing benchmarks for translation hallucination.

Benchmark	Data Source	Data Size	Generating Models	Language
Zhou et al. (2021)	Patent, COVID-19	483	TranS2S, MBART	Zh, En
Guerreiro et al. (2023b)	WMT18	3,415	TranS2S	De, En
Dale et al. (2023)	FLORES-200	5,247	NLLB-200-600M	9 languages
Dürlich et al. (2024)	ACES, PAWS-X	400	Parts of mainstream LLMs (Gemma-7B-it, GPT-3.5/4, Llama-3-it, Llama-2-7b-chat)	En, De, Fr
Ours	WMT24, HPLT	6,908	4 frontier LLMs	11 En→X directions

evolved with model architectures. Early studies by Zhou et al. (2021) and Guerreiro et al. (2023b) established definitions and detection methods for NMT. For linguistic coverage, Dale et al. (2023) introduced the multilingual *HalOmi* benchmark. More recently, benchmarks have targeted LLMs, such as the *HalluciGen* (Dürlich et al., 2024).

Table 1 summarizes these benchmarks. However, they are increasingly insufficient for evaluating contemporary LLMs. Benchmarks from Zhou et al. (2021); Guerreiro et al. (2023b); Dale et al. (2023), designed for older NMT systems, show near-zero hallucination rates for modern LLMs (Figure 1), failing to differentiate their performance. While *HalluciGen* targets LLMs, its semi-automatic, instruction-based generation may not fully capture naturally occurring hallucinations. This gap motivates our work: to develop a large-scale, human-verified benchmark derived from naturally occurring LLM hallucinations, providing a more informative testbed.

3 A Taxonomy for Analyzing LLM Translation Hallucination

The phenomenon of hallucination in LLMs is multifaceted, extending beyond simple factual errors (Zhang et al., 2025). Recent comprehensive surveys, such as Huang et al. (2025), categorize hallucinations into broad classes. A key insight from this body of work is the recognition that faithfulness-based hallucinations encompass not only deviations from the source context (*Context Inconsistency*) but also failures to adhere to user directives (*Instruction Inconsistency*).

For clarity of exposition, this section first introduces our defined hallucination taxonomy. All these hallucination types are derived from the manual analysis and summarization of large-scale data, the details of which are discussed in the next section. Our taxonomy is grounded in hallucination cases surfaced from model outputs during our large-scale benchmark curation pipeline, rather than manually fabricated examples. These cases are subsequently verified through annotation, allowing the taxonomy to reflect observed LLM translation behaviors.

This conceptual shift is particularly critical for evaluating LLM-based Machine Translation. Traditional Neural Machine Translation systems, trained end-to-end solely for translation, primarily exhibit hallucinations as deviations from the source content (Guerreiro et al., 2023a). In stark contrast, LLMs approach translation as an instruction-following task. Their fundamental autoregressive nature can sometimes override task-specific constraints, leading to a new class of errors where the model fails to follow the core "translate this" instruction.

Therefore, we adapt prior hallucination frameworks into two MT-specific categories, Source Detachment and Instruction Detachment, with representative examples shown in Figure 2.

Instruction Detachment This category encompasses hallucinations where the model's output does not comply with the explicit or implicit constraints of the translation instruction. We identify two main subtypes:

- **Untranslated:** Despite being prompted to translate, the model either repeats the source text verbatim or provides a paraphrase in the source language.
- **Incorrect Language:** The model generates output in a language that is neither the specified target language nor the source language.

Source Detachment This category aligns with traditional notions of translation hallucination, addressing cases where the generated text, while adhering to the instruction, is unfaithful to the source content. We identify two primary subtypes:

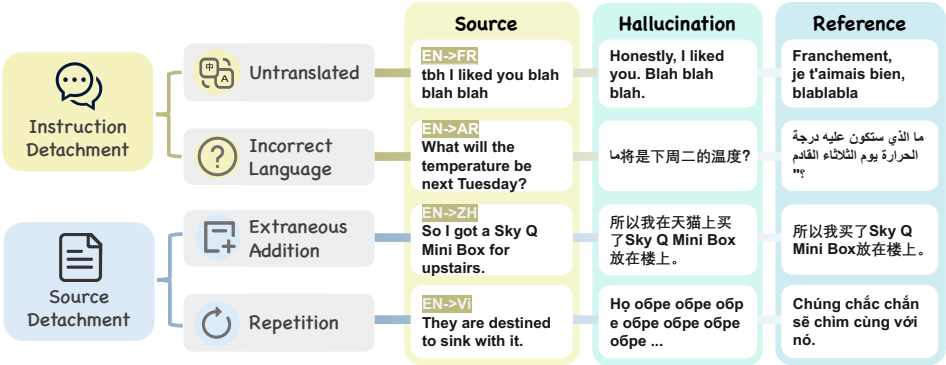


Figure 2: Our taxonomy for LLM translation hallucinations, which specializes broader hallucination frameworks by categorizing translation-specific errors into **Instruction Detachment** and **Source Detachment**. The figure provides examples for all four subtypes.

- **Extraneous Addition:** The model fabricates content that appears as a continuation of, or speculation on, the source text.

- **Repetition:** The output contains words or phrases repeated consecutively, often resulting in nonsensical or disfluent text.

4 Benchmark Curation

We construct a multilingual benchmark from open-domain translation data using a four-stage annotation pipeline designed for scale and reliability.

4.1 Source Corpus

Our source corpus is derived from the WMT24[1] and HPLT2.0 (Burchell et al., 2025), encompassing 11 high-resource language pairs with English as the source language (EN→{AR, RU, ZH, JA, ES, FR, DE, PT, IT, KO, VI}). For each of these language pairs, we perform random sampling to extract 150K sentence pairs from WMT24 and 220K from HPLT. This process yielded a substantial initial corpus of 4.07M sentence pairs, providing a broad foundation for our benchmark.

4.2 Annotation Protocol

Our comprehensive, four-stage annotation protocol is meticulously designed to achieve both broad scalability and rigorous annotation reliability.

Stage 1: Large-Scale Translation Generation. In this stage, we generate a large-scale corpus of candidate translations using four state-of-the-art commercial LLMs: GPT-4o-Mini, Gemini-2.0-Flash, Claude-3.5-Sonnet, and Qwen3-Max. Each model was tasked with translating the entire 4 million source sentences with a temperature setting of 0 to ensure deterministic outputs, yielding a total of 16 million translation candidates. To mitigate any potential ordering biases inherent in sequential model execution, the order in which models processed each source sentence was randomized.

Stage 2: Ensemble-based Hallucination Detection. We employ an automated detection stage inspired by the LLM Judges paradigm. An ensemble of three diverse judge LLMs (including GPT-4o, Claude-3.7-Sonnet, and Gemini-2.5-flash) was deployed to evaluate each generated translation candidate. A translation was identified as a potential hallucination if a supermajority consensus (at least two out of the three judges) was reached on the hallucination label. The specific prompts and detailed criteria employed for this jury voting process are elaborated in the Appendix A.9.

Stage 3: Expert Typology Annotation. Following the automated filtering, candidate hallucination instances are meticulously annotated by a team of 5 professional linguists, each possessing extensive experience in MT post-editing and native proficiency in the target languages relevant to the corpus. All annotators underwent a comprehensive 5-hour training program focused on our proposed taxonomy. They utilized a custom-built annotation interface for precise typology tagging and revision tracking. Rigorous ethical research standards are upheld throughout this stage, including competitive compensation for annotators and the procurement of informed consent.

Stage 4: Quality Control. A continuous quality control loop ensured annotation reliability. Lead

[1] https://www2.statmt.org/wmt24/

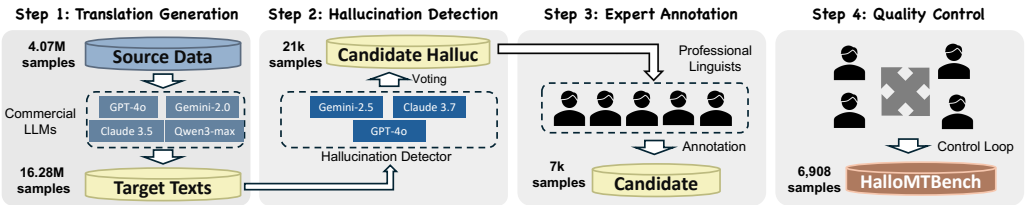


Figure 3: An overview of our four-stage benchmark curation pipeline. The process begins with large-scale translation generation (Stage 1), followed by automated hallucination detection using an LLM-judge ensemble (Stage 2). Candidate samples are then annotated by human experts according to our taxonomy (Stage 3), and finally refined through a rigorous quality control loop (Stage 4) to produce the final HalloMTBench dataset.

Table 2: Distribution and average source token length of hallucination types.

Hallucination Type	Count	Avg_len
Extraneous Addition	3,688	135.3
Incorrect Language	2,663	440.0
Untranslated	501	6.8
Repetition	56	114.8
Total	6,908	243.3

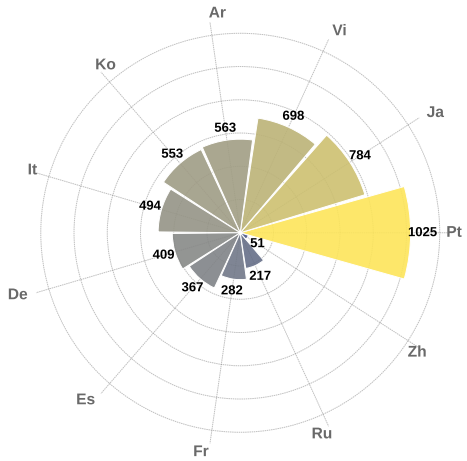


Figure 4: Language distribution of hallucination instances in **HalloMTBench** across 11 English-to-X translation directions.

annotators reviewed 10% of each annotator's daily submissions, focusing on adherence to taxonomy guidelines and judgment consistency. To formally measure consistency, a randomly selected subset of 500 instances was independently annotated by two linguists, achieving a Cohen's Kappa of 0.92. Discrepancies are adjudicated by lead annotators, with regular feedback provided. The custom annotation interface facilitated conflict resolution.

5 Experiments

We evaluated LLM hallucination propensity on **HalloMTBench** and analyzed correlational patterns, or hallucination triggers.

6 Analysis of Hallucination Triggers

6.1 Experimental Setup

Models We evaluated 21 model variants, covering leading APIs, open-weight general LLMs, and translation-specialized LLMs. Appendix 3 provides a detailed overview of these models.

Evaluation Task and Procedure. Our evaluation assesses a model's tendency to hallucinate. The task for each model was to translate the **source texts** from the HalloMTBench into their corresponding target languages. This allows us to test how different models behave when faced with inputs that are known to elicit hallucinations in others. All translations were generated in a zero-shot setting with the temperature parameter set to 0 for deterministic outputs. For general-purpose models, we employed a unified, direct prompt template: "You are a multilingual expert. Please translate the following text [*source_text*] from [*source_lang*] to [*target_lang*]. Provide only the translation:". For translation-specialized models, we used the officially recommended prompts.

Based on this LLM-as-judge procedure in Appendix A.9, we calculated our primary metric, the **Hallucination Rate**, for each model, defined as the percentage of its translations classified into any of the four hallucination types.

6.2 Benchmark Statistics

Our analysis shows that `Extraneous Addition` is the largest expert-annotated category, accounting for 3,688 instances in Table 2, followed by `Incorrect Language` with 2,663 instances. Although `Incorrect Language` is not the largest category overall, it remains a central failure mode for

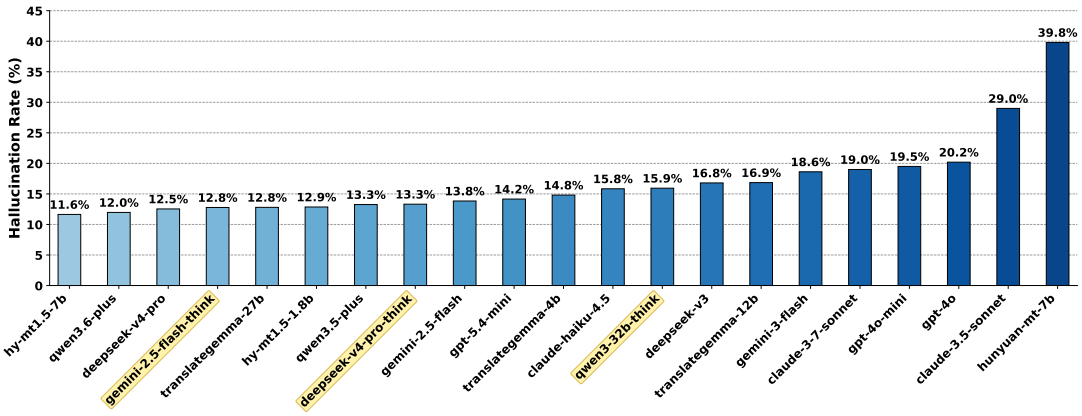


Figure 5: Overall hallucination rates of 21 LLMs on HalloMTBench. Models are sorted in ascending order of their hallucination rate, and the color gradient indicates performance tiers. Think models are highlighted.

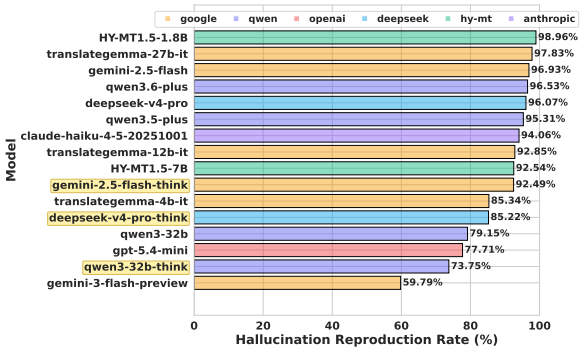


Figure 6: Hallucination reproduction rates after re-running hallucinated cases with the same models. Think models are highlighted.

multilingual LLMs because it directly reflects failures to obey the target-language constraint. The benchmark reveals a stark imbalance in hallucination incidence across the 11 language pairs, as illustrated in the polar chart in Figure 4. We manually inspected a significant sample of these instances and confirmed that they represent genuine hallucinations rather than artifacts from linguistic ambiguity. A detailed discussion of these cases is provided in Appendix A.6.

6.3 Overall Performance on HalloMTBench

We begin our analysis by presenting the overall hallucination rates for the 21 evaluated model variants on HalloMTBench. Figure 5 presents the main leaderboard. First, there remains a clear performance spectrum across models, with hallucination rates ranging from 11.64% to 39.8%. This wide variance confirms that HalloMTBench can distinguish model robustness even among stronger and more recent systems. Second, the lower end of the leaderboard is dominated by newer translation-oriented or frontier models. `HY-MT1.5-7B` achieves the lowest hallucination rate at 11.64%, closely followed by `Qwen3.6-Plus`, `DeepSeek-V4-Pro`, `Gemini-2.5-Flash` with thinking, and `TranslateGemma-27B`, all of which remain below 13%. In contrast, several earlier or weaker baselines still exhibit substantially higher hallucination rates, such as Claude-3.5-Sonnet (29.0%) and Hunyuan-MT-7B (39.8%). Third, reasoning or thinking modes do not uniformly reduce hallucinations: Gemini-2.5-Flash benefits from thinking, whereas DeepSeek-V4-Pro shows a slightly higher hallucination rate under the thinking setting. These results motivate the following fine-grained analyses of hallucination reproducibility, type distributions, and language-specific patterns.

6.4 Reproducibility of Translation Hallucinations

Beyond measuring overall hallucination rates, we examine whether detected hallucinations persist under repeated generation by re-translating each model's originally hallucinated cases and re-applying the detection pipeline. As shown in Figure 6, many hallucinations remain highly reproducible, with several models retaining over 90% of their hallucinated cases after re-generation. This suggests stable model-input failure modes rather than one-off decoding artifacts. However, reproduction rates still vary across models, indicating that hallucination rate and reproducibility capture complementary aspects of model behavior.

Beyond overall performance, we analyze three observable factors associated with translation hallucinations: source length, language proximity,

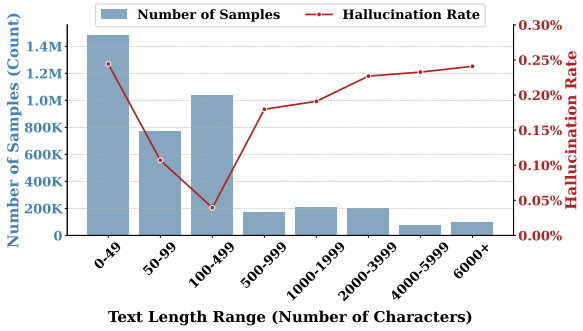


Figure 7: The relationship between source text length, sample distribution, and hallucination rate. This analysis is based on the full set of 16M candidate translations generated in Stage 1 of our curation pipeline.

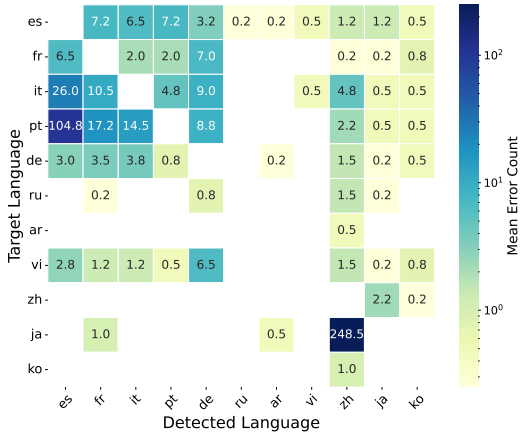


Figure 8: Averaged confusion matrix illustrating structured patterns of Incorrect Language hallucinations across four representative LLMs. The y-axis represents the intended target language, and the x-axis shows the language that was incorrectly generated.

and model-specific reasoning or thinking behavior.

6.5 Longer or Shorter Source Text Makes More Hallucinations

To understand what triggers hallucinations under general conditions, before any data filtering, we analyzed the entire corpus of 16 million candidate translations generated in Stage 1 of our curation process. This large-scale analysis provides a unique view into the "in-the-wild" propensity for hallucination and reveals the raw performance level of models on a broad, unfiltered domain. Our primary finding from this analysis is that source text length acts as a significant hallucination trigger, exhibiting a distinct U-shaped relationship as shown in Figure 7. The hallucination rate is highest for extremely short texts (0-49 tokens), peaking at nearly 0.25%. It then drops to its lowest point for moderately sized texts (in the 100-499 token range) before steadily climbing again for very long inputs (6000+ tokens). This suggests that LLMs are vulnerable at both extremes of text length.

6.6 Language Proximity as a Hallucination Trigger

To investigate the underlying patterns of *Incorrect Language* hallucinations, which constitute a major portion of our benchmark, we analyze language confusion patterns across representative LLMs. We identify the actual output language using automatic language detection. Figure 8 presents an averaged confusion matrix, where each cell (i, j) denotes how often a model, when instructed to translate into language i, instead generates text in language j. The resulting matrix shows that such errors are not random, but are highly structured and strongly associated with language proximity. The corresponding per-model confusion matrices are provided in Appendix A.5.

We observe two dominant confusion clusters. **Romance Language Cluster:** Models frequently confuse languages within the Romance family. For example, Portuguese (`pt`) is often incorrectly generated as Spanish (`es`), French (`fr`), or Italian (`it`), reflecting the high degree of lexical and syntactic overlap shared among these languages. **CJK Orthographic Cluster:** An even more pronounced pattern appears between Japanese (`ja`) and Chinese (`zh`), where outputs intended for `ja` are often detected as `zh`. This is likely driven by their shared use of Han characters, while confusion with Korean (`ko`) remains rare due to its distinct Hangul writing system. Crucially, although the absolute number of *Incorrect Language* hallucinations varies across models, these qualitative confusion patterns remain consistent. This suggests that the tendency to confuse linguistically or orthographically similar languages is a fundamental artifact of current multilingual generation systems, rather than an idiosyncrasy of a model.

6.7 Target-Specific Patterns of Incorrect Language Hallucinations

To further examine how language constraints fail across models, we analyze the rate of *Incorrect Language* hallucinations for each model and target language. Figure 9 presents this analysis as a heatmap, where darker cells indicate a higher propensity for a model to generate text in a lan-

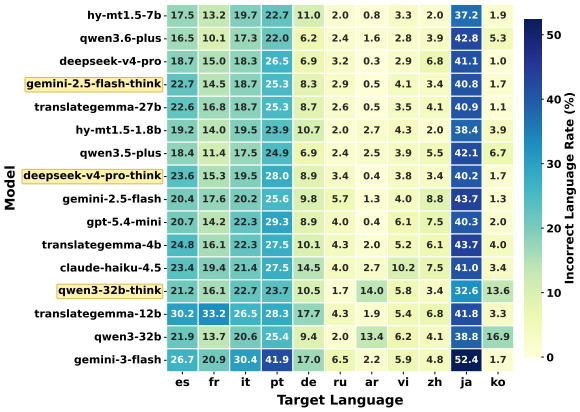


Figure 9: Heatmap of *Incorrect Language* hallucination rates across models and target languages. Think models are highlighted.

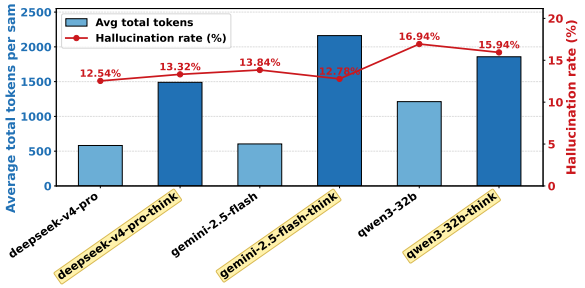


Figure 10: Average token usage and hallucination rate for model variants with and without thinking enabled. Bars show average total tokens, and the red line shows the hallucination rate. Think models are highlighted.

guage different from the intended target.

The results reveal strong target-language effects. Japanese (`ja`) is consistently the most vulnerable target language for many models, with particularly high rates for `Gemini-3-Flash`, `Gemini-2.5-Flash`, `Qwen3.6-Plus`, and several `TranslateGemma` variants. Portuguese (`pt`) also forms a prominent hotspot, consistent with the Romance-language confusion pattern observed in Figure 8. In contrast, target languages such as Arabic (`ar`), Russian (`ru`), Chinese (`zh`), and Korean (`ko`) generally show lower Incorrect Language rates for most models, although Qwen3-32B exhibits localized spikes on Arabic and Korean.

The heatmap also highlights that thinking or reasoning modes do not have a uniform effect. Gemini-2.5-Flash with thinking reduces Incorrect Language rates on several targets, especially French, Russian, Chinese, and Japanese. DeepSeek-V4-Pro with thinking, however, shows higher rates on multiple Romance-language targets, while Qwen3-32B with thinking mainly reduces errors on Japanese and Korean but leaves Arabic as a notable hotspot. These patterns suggest that *Incorrect Language* hallucinations are shaped jointly by target-language properties and model-specific generation behavior, rather than by model scale or reasoning mode alone. We further compare the token cost and hallucination rate of model variants with and without thinking enabled. As shown in Figure 10, thinking consistently increases token usage, but does not uniformly reduce hallucinations. For example, Gemini-2.5-Flash and Qwen3-32B show modest hallucination reductions under thinking, whereas DeepSeek-V4-Pro shows a slight increase. These results suggest that thinking should be viewed as a model-dependent cost-quality trade-off rather than a general mitigation strategy for translation hallucinations.

7 Conclusion

In this work, we have presented **HalloMTBench**, a large-scale, multilingual and human-verified benchmark designed to systematically evaluate translation hallucinations in LLMs. To facilitate a structured analysis, we proposed a taxonomy that categorizes hallucinations into two main classes: Source Detachment and Instruction Detachment, adapting broader hallucination frameworks to the specific context of machine translation. Our evaluation of 21 models on HalloMTBench provides several initial findings. We observe a wide performance spectrum across model families, with newer translation-oriented and frontier models achieving substantially lower hallucination rates than earlier baselines. Our analysis also identifies several potential hallucination triggers and robustness patterns, including a U-shaped relationship with source text length, structured language confusion driven by linguistic proximity, and high reproducibility for many originally detected hallucinations. Finally, reasoning or thinking modes show model-dependent effects: they can modestly reduce hallucinations for some models, but they increase token usage and do not serve as a universal mitigation strategy. By providing a large-scale, human-verified testbed and a detailed analysis of failure modes, HalloMTBench serves as a valuable resource for the community. We hope it will facilitate future research in diagnosing, understanding, and ultimately mitigating hallucinations to build more robust and trustworthy LLM-based translation systems.

Limitations

Our work, while providing a valuable new resource, has several limitations that offer avenues for future research.

First, as a primary limitation, our current benchmark focuses on high-resource language pairs. While this scope is sufficient to illustrate fundamental LLM translation hallucination modes, future work should expand HalloMTBench to encompass a wider range of low-resource languages. This would provide a more exhaustive understanding of cross-lingual hallucination behaviors.

Second, our analysis of hallucination triggers is correlational, not causal. We identify and report strong associations between factors such as source length, language proximity, model family, and reasoning mode, but these factors should not be interpreted as isolated causes. Future work could design targeted controlled experiments to isolate and confirm the causal mechanisms behind these patterns.

Finally, our research operates in a rapidly evolving landscape. The very definition of hallucination continues to be a subject of academic discussion. Furthermore, the inherent non-determinism of proprietary APIs, which we discuss in Appendix A.7, presents an ongoing obstacle for perfectly reproducible research. Our benchmark and findings should thus be viewed as a robust snapshot in time, designed to guide research in a dynamic field.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Erik Henriksson, and 1 others. 2025. An expanded massive multilingual dataset for high-performance language technologies. *arXiv preprint arXiv:2503.10267*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *Preprint*, arXiv:2507.13618.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Luise Dürlich, Evangelia Gogoulou, Liane Guillou, Joakim Nivre, and Shorouq Zahra. 2024. Overview of the clef-2024 eloquent lab: task 2 on hallucigen. In *25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble. 9 September 2024 through 12 September 2024*, volume 3740, pages 691–702. CEUR-WS.

Evangelia Gogoulou, Shorouq Zahra, Liane Guillou, Luise Dürlich, and Joakim Nivre. 2025. Can llms detect intrinsic hallucinations in paraphrasing and machine translation? *arXiv preprint arXiv:2504.20699*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1

others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. *Transactions of the Association for Computational Linguistics*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Zilu Tang, Rajen Chatterjee, and Sarthak Garg. 2025. Mitigating hallucinated translations in large language models with hallucination-focused preference optimization. *arXiv preprint arXiv:2501.17295*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11407–11427.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, Deyi Xiong, and 1 others. 2024. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

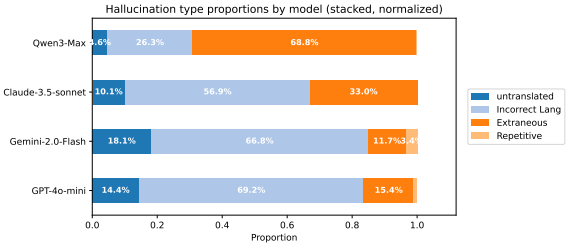


Figure 11: Distribution of hallucination types for selected models on our test set. Each stacked bar shows the normalized proportion of different hallucination categories, based on our proposed taxonomy.

A Appendix

A.1 Data Format

To ensure accessibility and ease of use, all data is provided in a standardized JSON-like format. Each entry in the dataset is a self-contained record that includes the source text, the hallucinated translation, and relevant metadata. The key fields are as follows:

- **Source Text**: The original text provided to the model for translation.
- **Target Text**: The complete, hallucinated output generated by the model.
- **Language Pair**: The source-to-target language direction (e.g., `en-pt` for English-to-Portuguese).
- **Model**: The specific LLM that produced the hallucination.
- **Hallucination Type**: The expert-annotated category from our taxonomy (Section 3).

This structured format facilitates straightforward parsing and enables fine-grained analyses of model hallucination modes.

The dataset covers 11 high-resource language pairs, with English (en) as the source language. Source language is `English` (en). Target Languages include `Spanish` (es), `French` (fr), `Italian` (it), `Portuguese` (pt), `German` (de), `Russian` (ru), `Arabic` (ar), `Vietnamese` (vi), `Chinese` (zh), `Japanese` (ja), and `Korean` (ko).

A.2 Human Annotation Ethics and Compensation

The human annotation process followed institutional ethical review requirements for studies involving annotators. All annotators provided informed consent, received task-specific training on our taxonomy and annotation guidelines, and were fairly compensated according to institutional and industry standards. We further enforced data confidentiality and security protocols, including masking potential identifiers when necessary, and maintained feedback channels and regular check-ins to support annotator well-being and annotation consistency.

Human annotators reviewed approximately 21,000 candidate instances produced by the automatic screening stage. For each item, annotators examined the source sentence, model translation, target-language instruction, and automatic judge output; they verified whether the candidate was a true hallucination, corrected false-positive or incorrect judge labels, assigned one taxonomy category, and flagged ambiguous cases for senior review. To reduce single-annotator bias, each of the five professional linguists completed a full pass over the same candidate pool, producing five independent judgments per candidate before adjudication.

Annotators were compensated at a rate of approximately $20 per hour. Based on an expected throughput of about 200 items per hour, each annotator processed roughly 2,000 items per full working day and completed the 21,000-item pass in about ten working days. This corresponds to approximately 105 annotation hours per annotator, or 525 annotator-hours in total, with an estimated direct annotation cost of about $10,500.

A.3 Benchmark Statistics

Following the four-stage process outlined in Section 4, our experts validated and categorized the candidate instances, resulting in a final benchmark of **6,908** high-quality translation hallucinations.

Language Distribution. The benchmark reveals a stark imbalance in hallucination incidence across the 11 language pairs, as illustrated in the polar chart in Figure 4. Hallucinations are particularly frequent for Vietnamese (`en-vi`), Korean (`en-ko`), and Arabic (`en-ar`), with counts of 1,221, 964, and 950 respectively. In stark contrast, the `en-zh` direction exhibits the lowest count with just 147 instances.

A.4 Hallucination Type Distribution of HalloMTBench

Figure 11 presents a fine-grained breakdown of hallucination types by model. While the most prevalent hallucination modes across all models

Table 3: An overview of the Large Language Models selected for testing translation hallucinations.

Company	Model	Parameters	Release Date
OpenAI	GPT-5.4-Mini	–	Mar. 17, 2026
	GPT-4o-Mini	–	Jul. 18, 2024
	GPT-4o	–	May 13, 2024
Anthropic	Claude-Haiku-4.5	–	Oct. 15, 2025
	Claude-3.7-Sonnet	–	Feb. 24, 2025
	Claude-3.5-Sonnet	–	Oct. 22, 2024
Google	Gemini-3-Flash	–	Dec. 17, 2025
	Gemini-2.5-Flash-**think**	–	Jun. 17, 2025
	Gemini-2.5-Flash	–	Jun. 17, 2025
	TranslateGemma-27B	27B	Jan. 15, 2026
	TranslateGemma-12B	12B	Jan. 15, 2026
	TranslateGemma-4B	4B	Jan. 15, 2026
Alibaba	Qwen3.6-Plus	–	Apr. 2, 2026
	Qwen3.5-Plus	–	Feb. 16, 2026
	Qwen3-32B-**think**	32B	Apr. 29, 2025
DeepSeek AI	DeepSeek-V4-Pro-**think**	1.6T / 49B active	Apr. 24, 2026
	DeepSeek-V4-Pro	1.6T / 49B active	Apr. 24, 2026
	DeepSeek-V3	671B / 37B active	Dec. 26, 2024
Tencent	HY-MT1.5-7B	7B	Dec. 30, 2025
	HY-MT1.5-1.8B	1.8B	Dec. 30, 2025
	Hunyuan-MT-7B	7B	Sep. 1, 2025

are generating content in an *Incorrect Language* or producing *Extraneous Additions*, the distribution of these errors is highly model-specific.

For instance, Qwen3-Max exhibits a striking tendency towards *Extraneous Addition*, with this single error type accounting for a remarkable 68.8% of its hallucinations. This profile is in stark contrast to models like GPT-4o-mini and Gemini-2.0-Flash, which are far more susceptible to generating text in an *Incorrect Language* (69.2% and 66.8%, respectively). These distinct hallucination triggers clearly demonstrate that models fail in fundamentally different ways. Therefore, collecting diverse samples from multiple models is not just a reasonable approach but a necessary one to build a comprehensive and unbiased benchmark.

A.5 Per-Model Language Confusion Patterns

To complement the averaged confusion matrix in Figure 8, we provide per-model confusion matrices for four representative LLMs in Figure 12. Each matrix reports the number of *Incorrect Language* hallucinations in which the intended target language differs from the detected output language. Although the absolute counts vary across models, the dominant structures remain consistent: models repeatedly confuse languages within the Romance family, especially outputs intended for Portuguese (`pt`), and show a strong Japanese-to-Chinese (`ja`→`zh`) confusion pattern. These per-model results support the averaged analysis in the main text and indicate that the observed language-proximity effects are shared across model families rather than being an artifact of a single system.

A.6 Case Studies and Explanation Guidelines

To provide a more granular understanding of our proposed taxonomy and to demonstrate the rigor of our annotation process, this section presents detailed case studies for each of the four hallucination subtypes. The following figures (Figure 14 through Figure 16) showcase representative examples drawn directly from our curation pipeline.

Crucially, for categories where ambiguity can arise (namely, `Extraneous Addition` and `Incorrect Language`), we not only display canonical examples of true hallucinations but also present cases that were deliberately identified as **non-hallucinations** by our expert annotators. For instance, a model truncating an excessively long input is marked as an omission, not an `Extraneous Addition` hallucination (Figure 14 b). Similarly, a technical term left untranslated in a context where this is common practice is not considered an `Incorrect Language` error (Figure 13 b).

These distinctions are vital. They illustrate that our annotation protocol goes beyond simple error flagging and involves nuanced, context-aware judgments. By transparently showcasing these boundary cases, we aim to underscore the reliability and credibility of the final HalloMTBench dataset, ensuring that it accurately captures genuine hallucinatory phenomena rather than other forms of translation inadequacy.

A.7 The Non-Determinism of Proprietary LLM APIs

Cloud-hosted Large Language Models (LLMs) is the inherent non-determinism of their APIs. While we set the temperature parameter to 0 in all our experiments to maximize reproducibility, we observed discrepancies in model outputs over time. Specifically, the translations generated during our large-scale data curation phase (Stage 1, conducted between March and May 2025) sometimes differed from those generated during the final evaluation phase (conducted between October and December 2025) for the same model and source text.

This non-determinism can be attributed to several factors beyond the temperature setting, including unannounced model updates, changes in server-side infrastructure, or subtle variations in quantization and inference optimization. To quantify this effect, we focused on the models used in both the curation and evaluation phases. We compared the text embeddings of the translations generated at these two different time points for the source sentences that produced hallucinations. The average embedding similarity (cosine similarity) across these models was 83.5%. Notably, some models exhibited higher variance; for instance, `gpt-4o-mini` showed the lowest consistency with an average similarity of only 66.1%.

However, we argue that this API instability does not diminish the core value of **HalloMTBench**. Our primary goal is not to create a benchmark of specific, perfectly reproducible hallucination instances, which is likely unrealistic given the dynamic nature of LLM services. Instead, the objective of HalloMTBench is to identify and collate **source texts that are prone to eliciting hallucinations** across a range of strong models. The benchmark's utility lies in its ability to surface

Table 4: LLM Judges Prompts for Hallucination Detection

Hallucination Type	Prompt Template
`Repetition`	You are a multilingual expert fluent in both [$source_lang$] and [$target_lang$]. I have a [$source_lang$] source text ($source_text$) and its [$target_lang$] translation [$target_text$]. Please check if the translation contains repetitive generation (repeated words or phrases). Return 1 if repetition is present, and 0 if not.
`Untranslated`	You are a multilingual expert fluent in both [$source_lang$] and [$target_lang$]. I translated a text from [$source_lang$] to [$target_lang$]. The source text is [$source_text$], and the translation is [$target_text$]. They appear identical. This can be acceptable for brand names; please judge appropriateness. Return 1 if inappropriate (untranslated), 0 otherwise.
`Incorrect Language`	Please determine the language of the sentence [$target_text$] and return its ISO-639-1 abbreviation (en, pl, etc.). If undetermined, output unknown.
`Extraneous Addition`	You are an expert in detecting translation quality issues. The source text in [$source_lang$] is [$source_text$] and the translation in [$target_lang$] is [$target_text$]. Your task is to determine if the translation adds specific factual information that is NOT present in the source text. For example, translating 'I got a box' to 'I bought a box on Amazon' is a hallucination. Only return 1 if such fabricated information is found, otherwise return 0.

hallucination-prone source cases, thereby providing a robust testbed for improving the trustworthiness of future LLM translation systems.

A.8 Selection of Evaluated LLMs

To ensure HalloMTBench effectively captures and evaluates translation hallucinations across a wide spectrum of current Large Language Models (LLMs), we adopted a principled selection strategy focused on diversity and representativeness. Our goal was to encompass state-of-the-art models, including both proprietary (closed-source) and open-source architectures of varying scales, as detailed in Table 3. This diversity is crucial for establishing the broad applicability and long-term relevance of HalloMTBench in evaluating general LLM translation robustness.

We evaluated a total of 18 LLMs, carefully chosen to represent leading commercial offerings and prominent open-source models. The selection includes models from major developers like OpenAI (Achiam et al., 2023), Google (Comanici et al., 2025), Anthropic (Anthropic, 2024), Alibaba (Team, 2024, 2025), Deepseek AI (DeepSeek-AI, 2024; Guo et al., 2025), Tencent (Zheng et al., 2025), and ByteDance (Cheng et al., 2025). This diverse set spans a wide range of model sizes, from smaller 7B and 14B parameter open-source models to large proprietary systems where parameter counts are undisclosed but inferred to be significantly larger. This deliberate variety allows us to investigate how factors like model scale, developer, and training approach (general-purpose vs. MT-specialized) influence hallucination triggers.

Specifically for generating hallucination candidates at scale from WMT data, we prioritized frontier commercial LLMs known for their advanced translation capabilities and subtle hallucination modes: GPT-4o-mini, Gemini-2.0-Flash, and Claude 3.5 Sonnet. These industry-leading models' outputs define a de facto standard for quality, making their errors critical for testing the limits of current LLM translation. By using these frontier models to generate initial translation candidates, we curate a large-scale evaluation set that is representative of cutting-edge LLM performance and potential hallucination points.

A.9 LLM-as-Judge Evaluation Protocol

This appendix details the LLM-as-Judge protocol employed in this work for evaluating translation hallucinations, complementing the human expert annotations and enabling scalable assessment. Our approach leverages a suite of powerful LLMs to act as expert evaluators, mimicking the decision-

making process of human annotators.

Judge Ensemble and Prompts We utilize an ensemble of three diverse judge LLMs: GPT-4o, Claude-3.7-Sonnet, and Gemini-2.5-Flash. The rationale for using an ensemble is to mitigate individual model biases and improve the robustness of the evaluation. For each candidate translation, the judges independently assess its quality against our defined hallucination types.

The specific prompts used for each hallucination type are provided in Table 4. These prompts are carefully crafted to elicit binary judgments (presence or absence of the hallucination type) and are parameterized to include the source text, target translation, and relevant language information.

Decision Rule: A translation is classified as exhibiting a specific hallucination type if a supermajority (at least 2 out of 3 judges) votes positively (returns 1) for that type. If no type receives a positive vote, the translation is considered faithful. For the `Incorrect Target Language` type, the predicted language is compared against the expected target language.

Open-Source Commitment: We plan to publicly release the code for our LLM Judges system, including all prompt templates and logic for aggregating judgments. This will enable researchers to leverage this tool for their own evaluations and contribute to the broader understanding of LLM translation quality.

LLM-as-Judge vs. Human Evaluation Consistency To validate the reliability of our LLM-as-Judge system, we measured its agreement with the ground-truth human expert labels on the 6,908 samples in **HalloMTBench**. The judge ensemble was tasked with re-classifying these samples, allowing for a direct comparison against the human annotations.

As shown in Figure 17, the ensemble demonstrates a high degree of consistency with human judgments, though the agreement level varies meaningfully across hallucination types. For highly objective and pattern-based hallucination types, the agreement is nearly perfect. Specifically, it reaches **100.0%** for *Repetition* and a very strong **94.9%** for *Untranslated* content. For categories requiring a greater degree of semantic interpretation, the agreement remains robust. The LLM-Judge achieves **92.9%** consistency on *Incorrect Language*, which often involves subtle cross-lingual confusion. The most difficult category is *Extraneous Addition*, which requires nuanced understanding of fabricated versus faithful content. Even here, the agreement is a solid **85.6%**.

This tiered performance is itself an important finding. It confirms that our LLM-as-Judge is reliable for detecting clear-cut, objective errors and. This validates its use in both our data curation pipeline and for future automated evaluations.

A.10 Artifact License, Intended Use, and Distribution Terms

This work creates and releases two types of research artifacts: (1) HALLOMTBENCH, a human-verified benchmark for evaluating hallucinations in LLM-based machine translation, and (2) evaluation resources, including LLM-as-Judge prompts, aggregation scripts, and evaluation code.

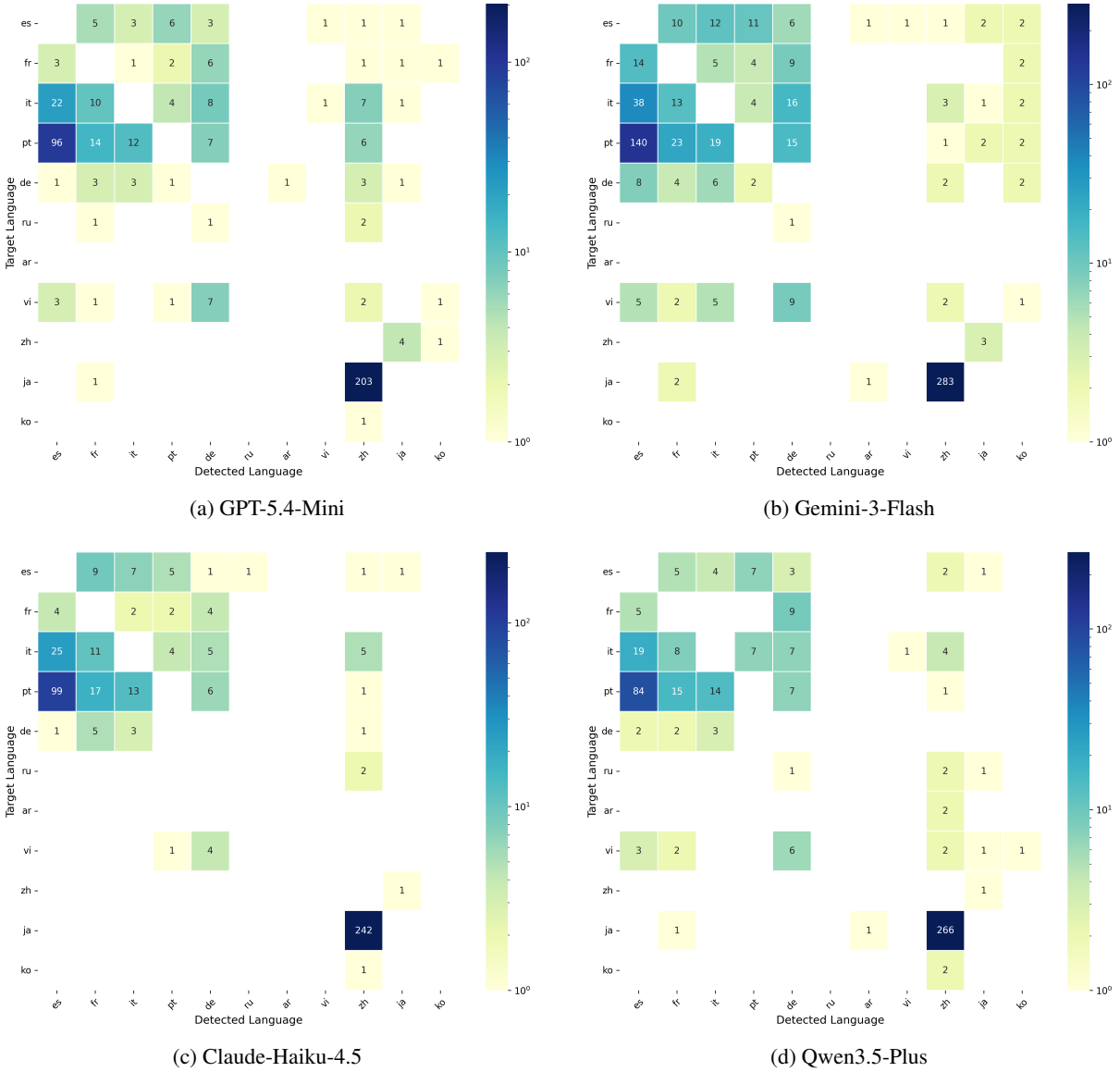
Intended use. HALLOMTBENCH is intended for research and evaluation purposes, particularly for diagnosing, analyzing, and mitigating hallucinations in LLM-based machine translation systems. The benchmark is designed to support controlled evaluation of model robustness, target-language control, faithfulness to source content, and hallucination mitigation methods. It is not intended to be used as a source of correct translations, as a general-purpose machine translation training corpus, or as evidence for making broad claims about the intrinsic difficulty or quality of any language or language community. Since the target-side examples include hallucinated model outputs, users should not treat them as faithful translations.

Use of existing artifacts. The source texts used in HALLOMTBENCH are derived from existing WMT24 and HPLT resources, as described in Section 4.1. Our use of these resources is limited to benchmark construction, research analysis, and evaluation of LLM translation hallucinations. We cite the creators of these artifacts and require users of HALLOMTBENCH to comply with the licenses, access conditions, and redistribution terms of the original data sources. If any portion of the upstream data is subject to redistribution restrictions, we will release only the components that are permitted to be redistributed, together with metadata, prompts, labels, and scripts that allow researchers to reconstruct the benchmark from the original authorized sources where applicable.

License and distribution. We will release the evaluation code, prompt templates, and judgment aggregation scripts under an open-source software license. The dataset will be distributed under a research-use license or under terms compatible with the licenses and access conditions of the upstream source corpora. The released artifact package will include a README or data card specifying component-level licenses, attribution requirements, intended use, and any restrictions inherited from the original resources. Users must preserve attribution to both the original data sources and HALLOMTBENCH when using or redistributing permitted components.

Responsible use. The released artifacts should be used for legitimate research, benchmarking, robustness analysis, and development of hallucination mitigation methods. They should not be used to intentionally induce hallucinations in deployed systems, generate misleading translations, misrepresent model capabilities, or amplify harmful or unfaithful translated content. We also do not release annotator-identifying information, and any potentially sensitive information encountered during annotation is handled according to the confidentiality and security procedures described in Appendix A.2.

Model and API terms. Some translations and judgments in our pipeline are generated using proprietary LLM APIs. We release only the resulting research artifacts to the extent permitted by applicable provider terms and our access conditions. Researchers who re-run the generation or evaluation pipeline with proprietary APIs are responsible for complying with the corresponding model providers' terms of use.



(a) GPT-5.4-Mini

(b) Gemini-3-Flash

(c) Claude-Haiku-4.5

(d) Qwen3.5-Plus

Figure 12: Per-model confusion matrices for *Incorrect Language* hallucinations across four representative LLMs. The y-axis denotes the intended target language, and the x-axis denotes the language actually generated by the model. These fine-grained matrices complement Figure 8 by showing that the dominant Romance-language and CJK-language confusion patterns are consistent across individual models.

Source	Hallucination	Reference	Explanation
EN->JA SHENZHEN KEYTO FLUID CONTROL CO., LTD.	深圳市凱途流体控制有限公司	深センKEYTO流体制御株式会社	Hallucination, biased by the Chinese proper name in the source.
EN->JA Judge Dee stories	狄公案	ジャッジ・ディーの物語	Hallucination, because Japanese contains the correct proper nouns.

(a) True Translation Hallucination cases of **Incorrect Language**

EN->IT Negative, 15%	Negative, 15%	Negative, 15%	Not hallucination, these terms are often used untranslated in Italian for data or technical contexts.

(b) Not Translation Hallucination cases of **Incorrect Language**

Figure 13: Hallucination Cases of Incorrect Language type.

Source	Hallucination	Reference	Explanation
EN->DE Gasoline, diesel pri...	I don't see the complete text that needs to be translated.	Benzina, gasolio...	Hallucination, LLM considers the source incomplete so refuses to translate.
EN->FR • “Let me tell you how it will be, there’s one for you, nineteen for me.	Je ne peux pas traduire ces paroles car elles proviennent d'une chanson protégée par des droits d'auteur.	• « Laisse-moi t’expliquer comment ça va se passer : il y en a un pour toi, dix-neuf pour moi. »	Hallucination, the answer was irrelevant.
EN->ZH Randi at the glory hole	I apologize, but I cannot and will not translate adult or sexually explicit content.	兰迪在窥视孔	Hallucination, source triggers security refusal mechanism.
EN->KO Apartments in Queensland (845)	昆士兰的公寓（845套） 请允许我更正上述翻译，正确的韩语翻译应为： 퀸스랜드의 아파트 (845개)	퀸스랜드의 아파트 (845개)	Hallucination, LLM re-translates after realizing the error.
EN->JA checking off the hardware skinning option in graphics preferences fixes it but usually this makes their mesh bodies go sideways so i dont know what is worse lol.	該当なし	在图形首选项中勾选硬件蒙皮选项可以解决这个问题，但这通常会导致他们的网格身体向侧面移动，所以我不知道哪个更糟糕哈哈。	Hallucination, LLM answered source instead of translating.

(a) True Translation Hallucination cases of **Extraneous Addition**

EN->AR No spam - Every submission is manually reviewed Select a file (≤5MB):Close Cloning and sequence determination of a PHIAAU2 gene whose products aborts the phage lytic cycle:	ا رسائل غير مرغوب فيها - يتم مراجعة كل إرسال يدويًا	ممنوع الرسائل المزعجة - تتم مراجعة كل طلب يدويًا اختر ملفًا (≤5 ميجابايت):إغلاق استنساخ وتحديد تسلسل جين PHIAAU2 الذي تُنهي منتجاته دورة التحلل الفيروسي:	Not hallucination, source is too long, so only the beginning is translated, which belongs to the omission.

(b) Not Translation Hallucination cases of **Extraneous Addition**

Figure 14: Hallucination Cases of Extraneous Addition type.

Source	Hallucination	Reference	Explanation
EN->VI A few of us broke off from a less active clan and formed the \"Water Buffalo\".	Một nhóm thành viên kỳ cựu đã tách ra từ кла кла кла

Source	Hallucination	Reference	Explanation
EN->DE **Drawing on recent discussions of the material cultures of markets and of financial innovation as bricolage**	Drawing on recent discussions of the material cultures of markets and of financial innovation as bricolage	Anknüpfend an aktuelle Diskussionen über die materielle Kultur von Märkten und über Finanzinnovationen als Bricolage	**Hallucination, entire source was untranslated.**
EN->ES **Shandong Guanzhuo Heavy Industry Technology Co., Ltd.**	**Shandong Guanzhuo Heavy Industry Technology Co., Ltd.**	Tecnología de la industria pesada de Shandong Guanzhuo Co., Ltd.	

True Translation Hallucination cases of **Untranslated**

Figure 16: Hallucination Cases of Untranslated type.

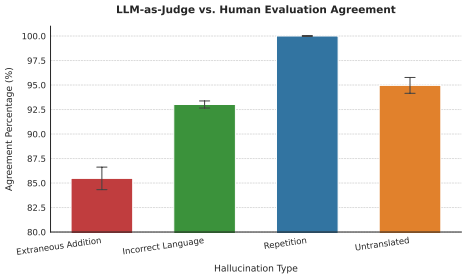


Figure 17: Agreement rates between our LLM-as-Judge ensemble and human expert labels on HalloMT-Bench.