Hallucinations in Large Multilingual Translation Models

Large-scale multilingual machine translation systems may generate hallucinated translations, which have the potential to severely undermine user trust and raise safety concerns. This work provides key insights regarding the prevalence, properties, and mitigation of hallucinations in massively multilingual machine translation systems.

Introduction

Recent advancements in large-scale multilingual machine translation have brought us closer to realizing a universal translation system, but these systems may still generate hallucinations, which can damage user trust and pose serious safety concerns.

Researchers have long recognized the problem of hallucinations in large-scale translation models, but most studies have been conducted on small bilingual models trained on a single English-centric high-resource language pair. This paper investigates hallucinations in two different classes of models: massively multilingual supervised models and generative LLMs.

This study analyzes two prevalent types of hallucinations in Neural Machine Translation(NMT) considered in the literature: hallucinations under perturbation and natural hallucinations. They also study a hybrid setup where other translation systems can be requested as fallback systems when an original system hallucinates.

Multilingual models predominantly struggle with hallucinations in low-resource language pairs and translating out of English; smaller distilled models can mitigate hallucinations by incorporating modeling choices that discourage them; ChatGPT produces hallucinations that are qualitatively different from those of conventional translation models.

2.1 Large Multilingual Language Models

Multilingual neural machine translation systems aim to translate directly with a single model for multiple language pairs without relying on any pivot language, and provide significant improvements over classic bilingual models.

As an alternative to supervised translation models, large language models (LLMs) can be pretrained on massive nonparallel corpora and can be prompted to solve arbitrary tasks. This has led to impressive results across a wide variety of NLP tasks.

2.2 Hallucinations in Machine Translation

Hallucinations in machine translation are rarer than in other natural language generation tasks, and are categorized into two types: hallucinations under perturbation and natural hallucinations.

Hallucinations under perturbation are translations that undergo significant negative shifts in quality due to perturbations in the source text. These translations occur naturally without any perturbation, and are categorized as largely fluent detached hallucinations or oscillatory hallucinations.

3.1 Models

This study focuses on two classes of models: conventional supervised multilingual NMT models and LLMs that can be prompted for translation. The supervised multilingual NMT models use the transformer-based M2M-100 family of models, and the LLMs use beam search with a beam size of 4.

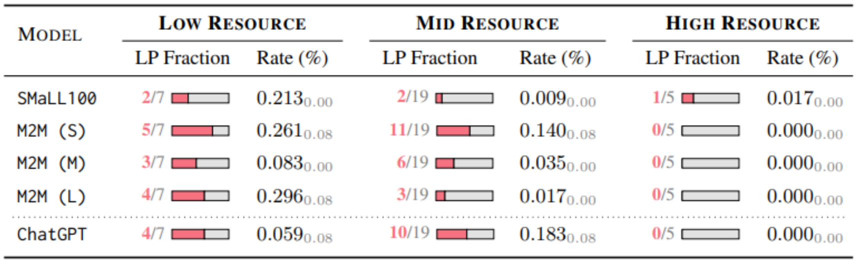
ChatGPT (gpt-3.5-turbo), a variant of GPT3.5, which has been fine-tuned with human feedback, to generate translations has also been used. It has been shown to achieve impressive results for multiple multilingual NLP tasks, including translation.

3.2 Datasets

Premier translation benchmarks like F LORES-101, WMT and TICO have been used. WMT is a high-quality multi-parallel dataset that consists of Wikipedia text in 101 languages and allows for the assessment of hallucinations across a vast range of translation directions.

3.3 Evaluation Metrics

spBLEU, COMET-22, CometKiwi, and LaBSE have been used to evaluate sentence-level translation. These have been successfully employed in prior research on detection of natural hallucinations.



MODEL	LOW RESOURCE		MID RESOURCE		HIGH RESOURCE	
	LP Fraction	Rate (%)	LP Fraction	Rate (%)	LP Fraction	Rate (%)
SMaLL100	2/7	$0.213_{0.00}$	2/19	$0.009_{0.00}$	1/5	$0.017_{0.00}$
M2M (S)	5/7	$0.261_{0.08}$	11/19	$0.140_{0.08}$	0/5	$0.000_{0.00}$
M2M (M)	3/7	$0.083_{0.00}$	6/19	$0.035_{0.00}$	0/5	$0.000_{0.00}$
M2M (L)	4/7	$0.296_{0.08}$	3/19	$0.017_{0.00}$	0/5	$0.000_{0.00}$
ChatGPT	4/7	$0.059_{0.08}$	10/19	$0.183_{0.08}$	0/5	$0.000_{0.00}$

Table 1: Fraction of languages for which models produces at least one hallucination under perturbation, and average hallucination rate (and median, in subscript) across all languages at each resource level.

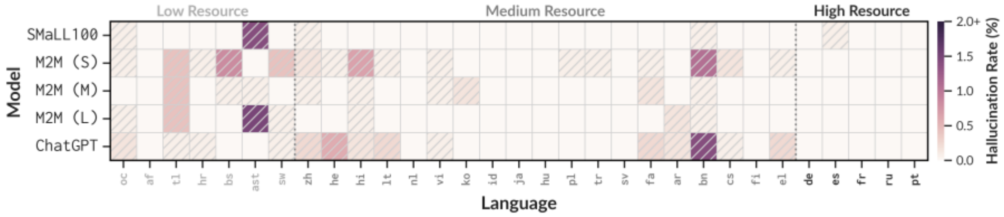


Figure 1: Heatmap of hallucination rates for each model in the languages considered. Pattern-filled cells indicate at least one hallucination under perturbation for a given model-language pair.

4.1 Evaluation Setting

Applying the same minimal perturbations used to construct the perturbed source sequences.

F LORES dataset is used to generate translations for 31 different language pairs, including bridge languages and additional low-resource languages.

The detection approach is inspired by previous works on hallucinations under perturbation. It is a simple 2-rule process that ensures that low-quality translations for unperturbed sources are not considered as candidates for hallucinations under perturbation. **Rules:** (i) a minimum threshold quality score for the original translations, and (ii) an extremely low maximum quality score for the perturbed translations. A model generates a hallucination under perturbation when both translations meet the thresholds.

This algorithm is extended to handle multiple models and language pairs by adapting rule (i). This ensures a fixed sample size across different language pairs, and allows us to effectively detect hallucinations under perturbation across multiple models in a multilingual scenario.

4.2 Results

Average hallucination rates decrease with increasing resource levels. SMaLL100 exhibits lower hallucination rates than its teacher model M2M (L), possibly because it was trained using uniform sampling across all language pairs to prevent bias towards higher resourced language pairs.

No correlation has been found between hallucinations under perturbation and the quality of original translations. Even minimal perturbations in the source text can cause significant shifts in translation quality.

ChatGPT exhibits different hallucination patterns from conventional translation models.

Table 1 shows that ChatGPT generates more hallucinations for mid-resource languages than for low-resource languages, and that these hallucinations are qualitatively different from those produced by traditional machine translation models.

Interestingly, most hallucinations can be reversed with further sampling from the model, suggesting that they may not necessarily indicate model defect.

5 Natural Hallucinations

They present a thorough analysis of natural hallucinations, including their different types, influence of translation direction, and prevalence of toxicity.

5.1 Evaluation Setting

Study analyzes massively multilingual translation models in three different evaluation scenarios, exploring more than 100 translation directions in the main text alone. It reports results for the first two setups using the F LORES dataset and WMT, and for the final setup using the TICO dataset.

Key findings from recent research on detection of hallucinations and focus on two main detectors have been integrated: ALTI+ for detached hallucinations and TNG for oscillatory hallucinations. ALTI+ and TNG have been validated on human-annotated hallucinations with perfect precision.

They rely on ALTI+, a model-based detector, for reliable detection of detached hallucinations. ChatGPT is excluded from our model selection to ensure consistency in our analysis.

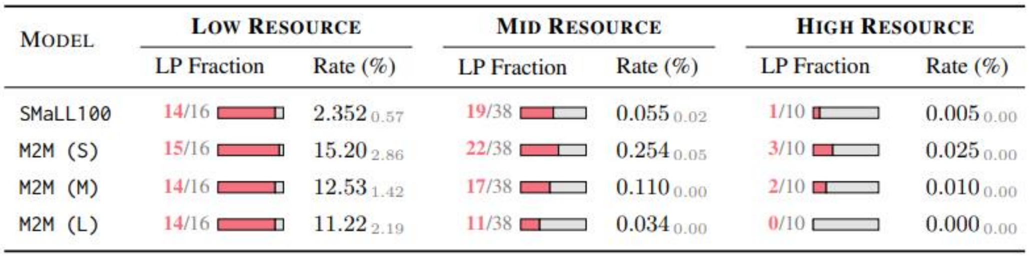
5.2 English-Centric Translation

Hallucinations are more frequent and distinct in low-resource languages than in mid- and high-resource languages, with detached hallucinations occurring more frequently. Furthermore, massive multilingual models exhibit average hallucination rates exceeding 10%.

SMaLL100 has the smallest number of parameters and shows remarkable hallucination rates across low- and mid-resource language pairs. Its improved rates may be attributed to its architectural decisions, which include a shallow 3-layer decoder and placing the target language code on the encoder side. SMaLL100's reduced hallucination rates do not necessarily imply superior translation quality compared to the other M2M models, as the correlation between M2M models' corpus-level COMET-22 scores and their respective hallucination rates is strong.

When translating out of English, models are significantly more prone to hallucinate than when translating into English. Furthermore, the translation direction can also influence the properties of hallucinations.

Toxic text in translations can emerge in the form of hallucinations, and is most prevalent in low-resource language pairs. These hallucinations are repeated across models for multiple unique source sentences, and are not necessarily reduced by scaling up the model size.



Model	Low Resource		Mid Resource		High Resource	
	LP Fraction	Rate (%)	LP Fraction	Rate (%)	LP Fraction	Rate (%)
SMaLL100	14/16	$2.352_{0.57}$	19/38	$0.055_{0.02}$	1/10	$0.005_{0.00}$
M2M (S)	15/16	$15.20_{2.86}$	22/38	$0.254_{0.05}$	3/10	$0.025_{0.00}$
M2M (M)	14/16	$12.53_{1.42}$	17/38	$0.110_{0.00}$	2/10	$0.010_{0.00}$
M2M (L)	14/16	$11.22_{2.19}$	11/38	$0.034_{0.00}$	0/10	$0.000_{0.00}$

Table 2: Fraction of LPs on the English-centric setup for which models produce at least one hallucination, and average hallucination rate (and median, in subscript) across all LPs at each resource level.

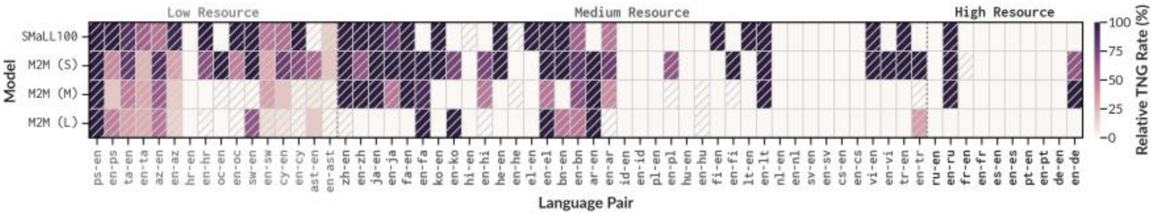


Figure 2: Heatmap of the percentage of hallucinations detected with TNG (oscillatory hallucinations) among all hallucinations. Pattern-filled cells indicate at least one natural hallucination for a given modelLP combination

5.3 Beyond English-Centric Translation

Focus is on directions that do not involve English, and find that models with less supervision during training exhibit extremely high hallucination rates.

5.4 Translation on Specialized Domains

They investigate hallucinations in medical domain data using the TICO dataset, and find that they are not exacerbated. This finding diverges from previous works that investigated hallucinations for specialized domain data, and suggests that the massive training set used in M2M models mitigates the impact of domain shift.

6 Mitigation of Hallucinations through Fallback Systems

The study explored the potential of using a hybrid setup to reduce hallucinations and enhance overall translation quality by leveraging an alternative system as a fallback when the primary original model produces hallucinations.

6.1 Employing models of the same family as fallback systems

The performance of same-family models is analyzed when employed as fallback systems for one another, and find that reversal rates are consistently higher for oscillatory hallucinations than for detached hallucinations. This finding emphasizes the close connection between detached hallucinations and training data.

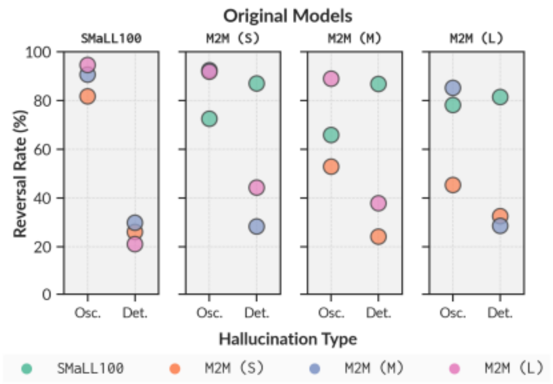


Figure 3: Reversal rates for oscillatory (Osc.) and detached (Det.) hallucinations when using models of the same family as fallback systems.

Scaling up within the model family is not an effective strategy for mitigating hallucinations.

Figure 3 shows that SMaLL100, a distilled M2M model, performs better at mitigating detached hallucinations than oscillatory hallucinations. This suggests that exploring alternative models with different architectures and trained on different data may yield more substantial improvements.

6.2 Employing external models as fallback systems

We will use a high-quality model from the NLLB family of multilingual NMT models to further mitigate hallucinations and improve translation quality.

Translation quality can be significantly improved with external fallback systems, particularly NLLB. NLLB outperforms ChatGPT as a fallback system for low- and mid-resource languages, and ChatGPT produces very few, if any, oscillatory hallucinations, slightly improving the rates obtained with NLLB.

7 Conclusion

This was an investigation on the phenomenon of hallucinations in massively multilingual translation models which found that hallucinations are prevalent across multiple translation directions across different resource levels and beyond English-centric translation, and that toxicity in hallucinations is a real problem.

Limitations

This study focuses on the M2M family of multilingual models, which includes the largest open-source multilingual NMT model.

Detection approaches inherit the limitations of the metrics that are leveraged in them, such as the BLEU metric.

This analysis showed that ChatGPT, a model that has demonstrated impressive capabilities for translation and other multilingual tasks, but we could not ensure that it was not trained on our evaluation sets.