Hallucinations in Large Multilingual Translation Models

Nuno M. Guerreiro[1,2,3] **Duarte M. Alves**[1,3] **Jonas Waldendorf**,[4]
Barry Haddow[4] **Alexandra Birch**[4] **Pierre Colombo**[5] **André F. T. Martins**[1,2,3]
[1]Instituto de Telecomunicações, Lisbon, Portugal [2]Unbabel, Lisbon, Portugal
[3]Instituto Superior Técnico, University of Lisbon, Portugal
[4]School of Informatics, University of Edinburgh, UK
[5]MICS, CentraleSupélec, Université Paris-Saclay, France
nuno.guerreiro@unbabel.com

Abstract

Hallucinated translations can severely undermine and raise safety issues when machine translation systems are deployed in the wild. Previous research on the topic focused on small bilingual models trained on high-resource languages, leaving a gap in our understanding of hallucinations in multilingual models across diverse translation scenarios. In this work, we fill this gap by conducting a comprehensive analysis—over 100 language pairs across various resource levels and going beyond English-centric directions—on both the M2M neural machine translation (NMT) models and GPT large language models (LLMs). Among several insights, we highlight that models struggle with hallucinations primarily in low-resource directions and when translating out of English, where, critically, they may reveal toxic patterns that can be traced back to the training data. We also find that LLMs produce qualitatively different hallucinations to those of NMT models. Finally, we show that hallucinations are hard to reverse by merely scaling models trained with the same data. However, employing more diverse models, trained on different data or with different procedures, as fallback systems can improve translation quality and virtually eliminate certain pathologies.

1 Introduction

Recent advancements in large-scale multilingual machine translation have brought us closer to realizing a universal translation system, capable of handling numerous languages and translation directions (Aharoni et al., 2019; Arivazhagan et al., 2019; Fan et al., 2020; Zhang et al., 2020; Wenzek et al., 2021; Goyal et al., 2022; NLLB Team et al., 2022). Concurrently, large language models (LLMs) have shown remarkable generalization to new tasks, including translation, where they are becoming increasingly stronger (Brown et al., 2020; Chowdhery et al., 2022; Hendy et al., 2023). Compared to traditional bilingual models, these systems can offer significant performance improvements and greatly simplify engineering efforts, as a single model can be used for all language pairs (Arivazhagan et al., 2019). As a result, they are an increasingly attractive choice for real-world applications. However, when deployed in the wild, these models may generate *hallucinations*: highly pathological translations detached from the source that can severely damage user trust and pose safety concerns (Perez et al., 2022).

The problem of hallucinations has long been recognized by researchers (Ji et al., 2022), and recent studies have contributed towards better understanding, detection and mitigation of these pathological translations. However, these studies have been conducted on *small bilingual* models (<100M parameters) trained on a *single English-centric high-resource* language pair (Raunak et al., 2021; Guerreiro et al., 2023b,a; Dale et al., 2023; Xu et al., 2023). This leaves a knowledge gap regarding the prevalence and properties of hallucinations in large-scale translation models across different translation directions, domains and data conditions.

In this work, we aim to fill this gap by investigating hallucinations on two different classes of models. The first and main class in our analysis is the *de facto* standard approach of massively multilingual supervised models: We use the M2M-100 family of multilingual NMT models (Fan et al., 2020), which includes the largest open-source multilingual NMT model with 12B parameters. The second class is the novel and promising approach of leveraging generative LLMs for translation. Contrary to conventional NMT models, these models are trained on massive amounts of

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1500–1517, 2023. https://doi.org/10.1162/tacl_a_00615
Action Editor: Stefan Riezler. Submission batch: 4/2023; Revision batch: 7/2023; Published 12/2023.

monolingual data in many languages, with a strong bias towards English, and do not require parallel data. In our analysis, we use ChatGPT and GPT-4, as thay have been shown to achieve high translation quality over a wide range of language pairs (Hendy et al., 2023; Peng et al., 2023).

We organize our study by analyzing the two prevalent types of hallucinations in NMT considered in the literature: hallucinations under perturbation and natural hallucinations (Lee et al., 2018; Raunak et al., 2021; Guerreiro et al., 2023b). Firstly, we study hallucinations under perturbation and evaluate whether these translation systems are robust to source-side artificial perturbations. While previous studies have found that these perturbations (e.g., spelling errors and capitalization mistakes) can reliably induce hallucinations (Lee et al., 2018; Raunak et al., 2021), it is not clear whether those conclusions hold for large multilingual models. Secondly, we comprehensively investigate natural hallucinations, and evaluate their properties in the outputs of the massively multilingual M2M models on a vast range of conditions, spanning from English-centric to non-English-centric language pairs, translation directions with little supervision, and specialized medical domain data where hallucinations can have devastating impact. Finally, we study a hybrid setup where other models can be requested as fallback systems when an original system hallucinates, with the aim of mitigating hallucinations and improving translation quality on-the-fly.

We provide several key insights on properties of hallucinations, including:

- models predominantly struggle with hallucinations in low-resource language pairs and translating out of English; critically, these hallucinations may contain toxic patterns that can be traced back to the training data;
- smaller distilled models can, surprisingly, hallucinate less than large-scale models; we hypothesize that this is due to modeling choices that discourage hallucinations (e.g., leveraging less potent shallow decoders that rely more on the encoder representations);
- LLMs produce hallucinations that are qualitatively different from those of conventional translation models, mostly consisting of off-target translations, overgeneration, and even failed attempts to translate;
- hallucinations are *sticky* and hard to reverse with models that share the same training data, whereas employing more diverse fallback systems can substantially improve overall translation quality and eliminate pathologies such as oscillatory hallucinations.

We release all our code and make available over a million translations in more than 100 translation directions to spur future research.[1]

2 Background

2.1 Large Multilingual Language Models

Massively multilingual neural machine translation has emerged as a powerful paradigm for building machine translation systems that can handle numerous languages (Akhbardeh et al., 2021; Wenzek et al., 2021; NLLB Team et al., 2022; Bapna et al., 2022; Chowdhery et al., 2022). These systems translate directly in multiple translation directions without relying on a pivot language.

The dominant strategy for achieving these systems is training large multilingual models on vast amounts of parallel data often obtained through a combination of data mining and data augmentation strategies, such as backtranslation (Sennrich et al., 2016; Edunov et al., 2018). Compared to traditional bilingual models, these systems bring significant improvements, particularly for low-resource and non-English-centric language pairs, as these benefit the most from multilingual transfer (Arivazhagan et al., 2019; Fan et al., 2020).

As an alternative, a novel strategy is to leverage LLMs. These systems are pretrained on massive nonparallel corpora and can be prompted to solve arbitrary tasks (Radford et al., 2019; Brown et al., 2020). In fact, this approach has led to impressive results across a wide variety of NLP tasks (Chowdhery et al., 2022; Zhang et al., 2022). Translation is no exception: LLMs can produce translations that are competitive with those of supervised translation models (Vilar et al., 2023; Peng et al., 2023; Hendy et al., 2023; Bawden and Yvon, 2023).

2.2 Hallucinations in Machine Translation

Hallucinations lie at the extreme end of translation pathologies and present a critical challenge in

[1] All resources are available in `https://github.com/deep-spin/lmt_hallucinations`.

machine translation, as they can compromise the safety and reliability of real-world applications.

Importantly, hallucinations in machine translation are unlike hallucinations in other natural language generation tasks (e.g., abstractive summarization) (Ji et al., 2022). Whereas, for these other tasks, models often produce hallucinated outputs (Falke et al., 2019; Cao et al., 2022; Manakul et al., 2023), hallucinations in machine translation, possibly due to the more closed-ended nature of the task, are substantially rarer and hard to observe in clean, unperturbed data. This has led several previous studies to examine their properties by creating artificial scenarios where hallucinations are more likely to occur (e.g., introducing perturbations in the source text (Lee et al., 2018) or noise in the training data (Raunak et al., 2021)). To distinguish these two scenarios, hallucinations in machine translation are categorized into two types (Raunak et al., 2021): *hallucinations under perturbation* and *natural hallucinations*.

Hallucinations under Perturbation. A model generates a hallucination under perturbation when it produces a significantly lower quality translation for a slightly perturbed input compared to the original input (Lee et al., 2018). Hallucinations under perturbation explicitly reveal the lack of robustness of translation systems to perturbations in the source by finding translations that undergo significant negative shifts in quality due to these changes.

Natural Hallucinations. These translations occur naturally, without any perturbation. In this work, we follow the taxonomy introduced in Raunak et al. (2021) and extended in Guerreiro et al. (2023b). Under this taxonomy, hallucinations are translations that contain content that is detached from the source, and are further categorized as *fluent detached hallucinations* or *oscillatory hallucinations*. The former refers to translations that bear minimal or no relation at all to the source, while the latter refers to inadequate translations that contain erroneous repetitions of words and phrases.

While recent research has contributed towards understanding, detection, and mitigation of hallucinations (Guerreiro et al., 2023a,b; Dale et al., 2023; Xu et al., 2023), the scope of these studies has been restricted to small models ($<$100M parameters) trained on a single, high-resource language pair. In this work, we expand upon this prior research by studying hallucinations in large multilingual models across various translation directions, domains, and data conditions, thereby addressing an important gap in the literature.

3 Experimental Suite

3.1 Models

For supervised multilingual NMT models, we use the transformer-based (Vaswani et al., 2017) M2M-100 family of models (Fan et al., 2020): `M2M (S)` with 418M parameters, `M2M (M)` with 1.2B parameters, and `M2M (L)` with 12B parameters. These models were trained on a many-to-many parallel dataset of 7.5B sentences, and support 100 languages and thousands of language pairs (LPs). We also evaluate `SMaLL100` (Mohammadshahi et al., 2022), a model with 330M parameters obtained via distillation of `M2M (L)`. `SMaLL100` was trained on a smaller training set obtained via uniform sampling across all language pairs to reduce the bias towards high-resource languages: only 100k parallel sentences from the M2M training data were used for each LP, for a total of 456M parallel sentences. For decoding, we run beam search with a beam size of 4.

As for the alternative strategy using LLMs, we use two GPT models: `ChatGPT` (`gpt-3.5-turbo`) and `GPT-4`.[2] These models are upgraded versions of GPT-3.5—a 175B GPT (Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020) LLM—that has been fine-tuned with human feedback in the style of InstructGPT (Ouyang et al., 2022). In particular, `ChatGPT` has been shown to achieve impressive results for multiple multilingual tasks, including translation (Kocmi and Federmann, 2023; Fu et al., 2023; Hendy et al., 2023). Crucially, GPT models, unlike the majority of existing LLMs, exhibit extensive and strong capabilities across multiple language pairs, a prerequisite for carrying out the research in this work. To generate translations, we use the zero-shot prompt template used in Hendy et al. (2023) and the default API parameters.

[2] `https://platform.openai.com/docs/models/`; we used `ChatGPT` (`gpt-3.5-turbo`) and `GPT-4` (`gpt-4`) in March, April, and June 2023.

3.2 Datasets

We chose datasets familiar to researchers and practitioners, ensuring no train/test overlap for the M2M models. To this end, we selected premier translation benchmarks: FLORES-101 (Goyal et al., 2022), TICO (Anastasopoulos et al., 2020), and WMT. FLORES-101 is a multi-parallel dataset that consists of Wikipedia text in 101 languages; we join the `dev` and `devtest` sets. TICO is a specialized medical-domain multilingual benchmark with COVID-19 related data, such as medical papers and news articles; we join the `dev` and `test` sets. Additionally, we use (i) WMT benchmarks from the M2M paper evaluation suite, as they were removed from the training data of M2M models;[3] and (ii) WMT 2022 benchmarks (Kocmi et al., 2022), as they were created after the cutoff date of training data of all tested models.

3.3 Evaluation Metrics

Our main lexical metric is spBLEU (Goyal et al., 2022), as it has been widely employed in works on massively multilingual translation (Fan et al., 2020; Wenzek et al., 2021; NLLB Team et al., 2022).[4] Moreover, we also adopt neural reference-based and reference-free COMET variants: COMET-22 and CometKiwi (Rei et al., 2022a,b). Lastly, we use the cross-lingual encoder LaBSE (Feng et al., 2022) to obtain sentence similarity scores, as these have been successfully used in prior research on detection of hallucinations (Guerreiro et al., 2023a; Dale et al., 2023).

4 Hallucinations under Perturbation

We start our analysis by focusing on artificially created hallucinations. We first provide an in-depth overview of our evaluation setting, focusing on the construction of the perturbed data and detection approach. Then, we present our results and analyze the properties of these hallucinations.

4.1 Evaluation Setting

Translation Directions. We use the FLORES dataset for these experiments, and focus specifically on translation out of English.[5] We select all M2M bridge languages, as well as additional low-resource languages that were underrepresented among bridge languages.[6] Overall, we generate translations for 31 different LP.[7] Additionally, we report results for the WMT 2022 benchmarks in Appendix A.3 to ensure evaluation on data that was released after the cutoff date of the training data of GPT models.[8]

Perturbations. We employ the same minimal perturbations used in Xu et al. (2023) to construct perturbed source sequences: misspelling errors, insertion of frequent tokens in the beginning of the sequence, and capitalization errors. Previous work has shown that these perturbations can trigger severe output errors (Lee et al., 2018; Karpukhin et al., 2019; Berard et al., 2019; Raunak et al., 2021). Additionally, we also experiment with the cascade approach in speech translation (Bentivogli et al., 2021), providing a real-world scenario where translation models have to deal with noisy inputs. In this approach, the audio (for consistency, we use the FLEURS dataset [Conneau et al., 2022], which consists of audio recordings of the sentences in the FLORES dataset) is first transcribed by an automatic speech recognition (ASR) system (we use a Whisper model [Radford et al., 2022][9]) and then translated by a translation system.

Detection. Our detection approach is inspired by previous work on hallucinations under perturbation (Lee et al., 2018; Raunak et al., 2021; Ferrando et al., 2022; Xu et al., 2023). The algorithm is a simple 2-rule process: we fix (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.

[3] These WMT benchmarks serve as our validation set for tuning thresholds as needed.

[4] Signature: `nrefs:1|case:mixed|eff:yes|tok:flores101|smooth:exp|version:2.3.1`.

[5] The training data—up to September 2021—of GPT models is not publicly available. As such, we cannot guarantee that they have not seen the data we use in this analysis.

[6] Fan et al. (2020) divide the 100 languages into 14 related groups (e.g., Romance, Slavic) and mine languages within a group against each other. Then they define 1-3 ''bridge'' languages per group (often those with most resources) and mine them against each other to connect the groups.

[7] `af ar ast bn cs cy de el es fa fi fr he hi hr hu id ja ko lt nl oc pl pt ru sv sw tl tr vi zh`.

[8] The results and trends on the WMT 2022 test sets follow largely those reported in the main text on the FLORES dataset.

[9] We used the `openai/whisper-base.en` model from the HuggingFace hub. Transcriptions with word error rate over 100 ($\sim$ 3% of samples) were discarded, as these represent significant, rather than minimal, perturbations.

Model	Low Resource		Mid Resource		High Resource	
	LP Fraction	Rate (%)	LP Fraction	Rate (%)	LP Fraction	Rate (%)
SMaLL100	4/7	$0.216_{0.06}$	8/19	$0.027_{0.00}$	1/5	$0.012_{0.00}$
M2M (S)	6/7	$0.392_{0.19}$	14/19	$0.172_{0.06}$	0/5	$0.000_{0.00}$
M2M (M)	5/7	$0.108_{0.06}$	10/19	$0.047_{0.06}$	0/5	$0.000_{0.00}$
M2M (L)	4/7	$0.327_{0.06}$	4/19	$0.020_{0.00}$	0/5	$0.000_{0.00}$
ChatGPT †	4/7	$0.082_{0.06}$	16/19	$0.202_{0.12}$	0/5	$0.000_{0.00}$
GPT-4 †	2/7	$0.019_{0.00}$	7/19	$0.057_{0.00}$	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) among candidate translations across all languages at each resource level. †GPT models may have been exposed to the test samples.

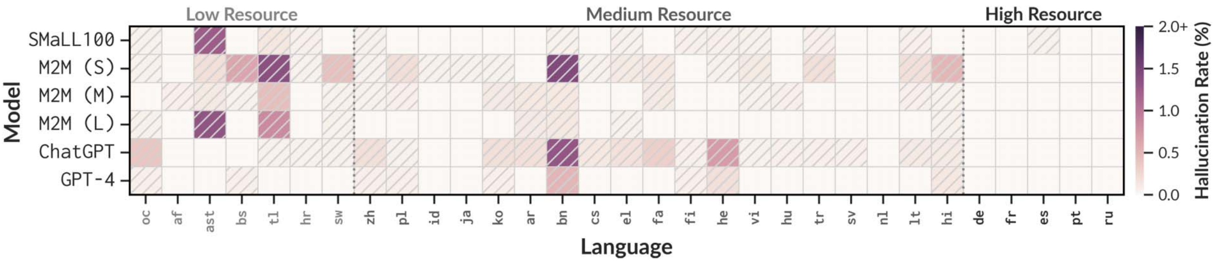


Figure 1: Hallucination rates among candidate translations for each model in the languages considered. Pattern-filled cells indicate at least one hallucination for a given model-language pair.

Crucially, rule (i) ensures that low-quality translations for unperturbed sources are not considered as candidates for hallucinations under perturbation.

We adapt rule (i) to ensure consistency across multiple models and LPs. We first obtain source sentences for which all models produce translations that meet a minimum quality threshold (spBLEU > 9), sort them by average quality across models, and select the top 20% as candidates. Finally, we apply rule (ii) and set the threshold to spBLEU < 3.[10] We selected these thresholds based on those used in previous works (Raunak et al., 2021; Ferrando et al., 2022; Xu et al., 2023). In Appendix A.1, we validate our detection method with human annotation on over 200 translations for 10 different language pairs.[11]

4.2 Results

Overall, we find that perturbations have the potential to trigger hallucinations, even in larger models. In what follows, we present several key insights.

Average hallucination rates for NMT models generally decrease with increasing resource levels. Table 1 shows that all NMT models exhibit lower hallucination rates as resource levels increase. This is expected and suggests that these models are better equipped to handle source-side perturbations for language pairs with more parallel data during training. In fact, hallucinations under perturbation for high-resource languages are almost non-existent. However, Figure 1 reveals variability across languages, and even within the models in the same family that share the same training data. For instance, when translating to Asturian (ast), M2M (L) and the distilled SMaLL100 show significantly higher hallucination rates than M2M (S). This suggests that hallucinations emerge in non-trivial ways unrelated to the training data.

[10] This approach ensures a fixed sample size across different LPs, and that the sentences for each LP are consistent across all models. The downside is that a system's hallucination rate depends on the systems it is being compared against.

[11] We found that over 85% of the detected hallucinations under perturbation were annotated as containing content detached from the source text.

SMaLL100 exhibits lower hallucination rates than its teacher model M2M (L). Recall that SMaLL100 was trained using uniform sampling across all language pairs to reduce bias towards high-resource languages. The results in Table 1 may reflect a positive outcome from this approach: despite being much smaller than M2M (L), SMaLL100 hallucinates less and for fewer directions on low- and mid-resource language pairs.

Hallucinations under perturbation are not correlated with the quality of original translations. The approach for detection of hallucinations under perturbation raises an interesting question: *Are the original source sentences for which models produce higher quality translations less likely to lead to hallucinations when perturbed?* Our analysis found a weak Pearson correlation[12] between hallucinations under perturbation and spBLEU scores for the original unperturbed sources across all models. This indicates that the perturbations that we introduce in the source can cause models to undergo significant shifts in translation quality.

LLMs exhibit different hallucination patterns from conventional NMT models. Contrary to NMT models, the GPT models generate more hallucinations for mid-resource languages than for low-resource languages (Table 1). When compared to all other models, the results show that ChatGPT produces more hallucinations for mid-resource languages, while GPT-4 exhibits fewer hallucinations in low-resource directions and across fewer languages. Additionally, hallucinations from LLMs are qualitatively different from those of other models: They often consist of off-target translations, overgeneration, or even failed attempts to translate (e.g., *''This is an English sentence, so there is no way to translate it to Vietnamese''*). This further demonstrates that translation errors, even critical ones, obtained with LLMs are different from those produced by NMT models (Vilar et al., 2023; Hendy et al., 2023).

Interestingly, we also found that almost all hallucinations can be reversed with additional sampling from the model. This aligns with findings in Guerreiro et al. (2023b) and Manakul et al. (2023): As with traditional NMT models, LLM hallucinations may not necessarily imply model defect or inability to generate adequate translations, but could just stem from ''bad luck'' during generation.

5 Natural Hallucinations

We now turn to investigating natural hallucinations.[13] We first provide an overview of our evaluation setting, focusing on the scenarios and detection approach. Then, we present a thorough analysis exploring various properties of hallucinations.

5.1 Evaluation Setting

Evaluation Scenarios. Analyzing multilingual translation models opens up several research scenarios that have not been explored in previous studies conducted on bilingual models. We will investigate three different multilingual scenarios, comprising over 100 translation directions.

We begin with an English-centric scenario, pairing 32 languages with English for a total of 64 translation directions.[14] Then, we study a non-English-centric scenario inspired by Fan et al. (2020), exploring 25 language pairs corresponding to real-world use cases of translation not involving English.[15] Finally, we examine hallucinations in medical data, where they can have severely compromising effects. We pair 9 languages[16] with English for a total of 18 directions. We report results for the first two setups using the FLORES dataset. For the final setup, we use the TICO dataset.

Detection. We integrate key findings from recent research on detection of hallucinations and employ two main detectors: ALTI+ (Ferrando et al., 2022) for detached hallucinations, and top n-gram (TNG) (Raunak et al., 2021, 2022; Guerreiro et al., 2023b) for oscillatory hallucinations.

ALTI+ assesses the relative contributions of source and target prefixes to model predictions. As hallucinations are translations detached from

[12] The point-biserial Pearson correlation between the hallucination assignments and the original sentence spBLEU scores is in the range $[-0.03, -0.01]$ for all models.

[13] From now on, we use the terms natural hallucinations (both detached and oscillatory hallucinations) and hallucinations interchangeably.

[14] ar ast az bn cs cy de el es fa fi fr he hi hr hu id ja ko lt nl oc pl ps pt ru sv sw ta tr vi zh.

[15] hi-bn it-fr de-hu it-de cs-sk nl-fr fr-sw ro-ru ro-uk de-hr hr-sr be-ru hr-hu hr-cs el-tr hr-sk nl-de af-zu ro-hu hi-mr ro-tr uk-ru ro-hy ar-fr ro-de.

[16] ar fr hi id ps pt ru sw zh.

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.

the source sequence, ALTI+ can be leveraged to detect them by identifying sentences with minimal source contribution. Notably, it faithfully reflects model behavior and explicitly signals model detachment from the source in any translation direction (Ferrando et al., 2022). This method has been successfully employed to detect hallucinated toxicity in a multilingual context in NLLB Team et al. (2022), and validated on human-annotated hallucinations in Dale et al. (2023), where it was shown that ALTI+ scores easily separate detached hallucinations from other translations.[17]

TNG, on the other hand, is a simple, lightweight black-box heuristic targeting oscillatory hallucinations. It compares the count of the top repeated translation n-gram to the count of the top repeated source n-gram, ensuring a minimum difference of t. This method has been validated on human-annotated hallucinations and found to identify oscillatory hallucinations with perfect precision (Guerreiro et al., 2023b). Following previous work, we use $n = 4$ and $t = 2$ (Raunak et al., 2021; Guerreiro et al., 2023b) and exclude translations that meet the minimum quality threshold from Section 4.1.[18]

Remark on Model Selection. We use ALTI+, a model-based detector, for reliable detection of detached hallucinations. Since we lack access to internal features from the GPT models, we exclude them from our model selection to ensure consistency in our analysis. Importantly, using alternative detectors could lead to misleading results and create discrepancies in our evaluation setup.

5.2 English-Centric Translation

We start by studying hallucinations on English-centric language pairs. We reveal key insights on how properties of hallucinations change across resource levels, models and translation directions.

Hallucinations in low-resource language pairs are not only more frequent, but also distinct. Table 2 shows that hallucinations occur frequently for low-resource directions, with all M2M models exhibiting average hallucination rates over 10%. Moreover, all models generate hallucinations for almost all low-resource language pairs. Regarding the type of hallucinations, Figure 2 shows that in low-resource directions, in contrast to mid- and high-resource ones, oscillatory hallucinations are less prevalent, while detached hallucinations occur more frequently. These findings suggest that, although massive multilingual models have significantly improved translation quality for low-resource languages, there is considerable room for improvement, and also highlight potential safety issues arising from translations in these directions.

`SMaLL100` consistently relies more on the source than other models. Despite being the smallest model, `SMaLL100` shows remarkable hallucination rates across low- and mid-resource directions, hallucinating significantly less than its larger counterparts in low-resource ones. We hypothesize that these improved rates may be attributed not only to the uniform sampling of language

[17] We followed the recommendations in Guerreiro et al. (2023b) and set model-based ALTI+ thresholds based on validation data where the models are expected to perform well. Specifically, we obtained the lowest 0.02%—in line with natural hallucination rates reported in the literature (Raunak et al., 2022)—of the ALTI+ score distributions for high-resource WMT benchmarks. Additionally, to ensure further trustworthy, high-precision measurements, we excluded detected candidates with LaBSE or CometKiwi scores—as these have been also been validated for detection of human-annotated detached hallucinations (Dale et al., 2023; Guerreiro et al., 2023a)—exceeding the top 10% of scores on translations from the same WMT benchmarks.

[18] Note that oscillatory hallucinations can be simultaneously detected with ALTI+ and TNG.

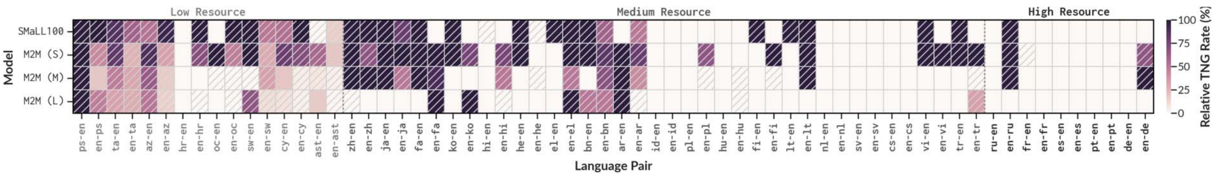


Figure 2: Heatmap of the rate of hallucinations detected with TNG (oscillatory hallucinations) among all hallucinations. Patterned cells indicate at least one natural hallucination for a given model-LP pair.

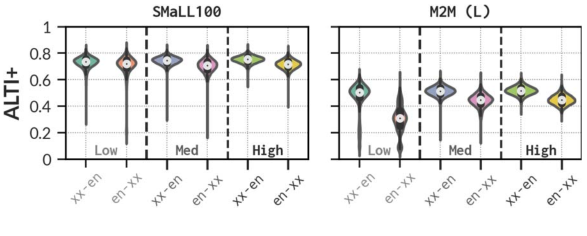


Figure 3: Distribution of SMaLL100 and M2M (L) ALTI+ scores for en-xx and xx-en directions.

LPs	SMaLL100	M2M (S)	M2M (M)	M2M (L)
xx-en	0.221	1.756	2.290	2.483
en-xx	1.022	6.152	4.110	3.169

Table 3: Average hallucination rates (%) across all LPs at each direction (into or out of English).

pairs during training, but also to its architecture. While SMaLL100 shares a 12-layer encoder with the other models to obtain source representations, it diverges by employing a shallow 3-layer decoder—instead of a 12-layer decoder—and placing the target language code on the encoder side. This design may encourage greater reliance on the more complex encoder representations, reducing detachment from the source. In fact, distinct patterns in ALTI+ scores support this hypothesis: SMaLL100 has higher scores and similar patterns across all resource levels (see Figure 3). In contrast, the M2M models tend to rely less on the source, especially in low-resource en-xx LPs. Importantly, however, SMaLL100's lower hallucination rates do not necessarily imply superior translation quality compared to the M2M models: We found a strong correlation between M2M models' COMET-22 scores and their respective hallucination rates for low-resource LPs, whereas, contrastingly, the correlation is weak for SMaLL100.[19] This suggests that, despite relying more on the source, SMaLL100's translations may not necessarily be of higher quality than those of the M2M models.

Scaling up models within the same family reduces hallucination rates. Table 2 shows that increasing the size of the M2M models results in consistent reductions in hallucination rates. Relative improvements are more pronounced for mid- and high-resource language pairs, with M2M (L) exhibiting fewer hallucinations and hallucinating for fewer languages than all other models.

Hallucinations are more frequent when translating out of English. Table 3 demonstrates that models are consistently more prone to hallucinate when translating out of English. In fact, models tend to detach more from the source text in these directions. This is evidenced by ALTI+ source contributions (see Figure 3) being lower across all en-xx language pairs compared to translating into English, which aligns with observations in Ferrando et al. (2022). Interestingly, we also discovered that the translation direction can influence the properties of hallucinations: (i) over 90% of off-target hallucinations occur when translating out of English, and (ii) nearly all hallucinations into English for mid- and high-resource language pairs are oscillatory (see Figure 2).

Toxic hallucinations can be traced back to the training data. To assess the prevalence of toxic text in detected hallucinations, we utilized the toxicity wordlists provided by NLLB Team et al. (2022). We found that toxic text predominantly appears in translations out of English and almost exclusively for low-resource directions. For instance, over 1 in 8 hallucinations in Tamil contain toxic text. Interestingly, these hallucinations exhibit high lexical overlap among them and are repeated across models for multiple unique source sentences. Moreover, they are not necessarily reduced by scaling up the model size. These observations suggest that these hallucinations are likely to be traced back to the training data, aligning with observations in Raunak et al. (2021) and Guerreiro et al. (2023b). In fact, upon inspecting

[19] Pearson correlation scores: SMaLL100 (−0.39), M2M (S) (−0.82), M2M (M) (−0.80), M2M (L) (−0.85).

Model	Low Resource		Mid Resource		High Resource	
	LP Fraction	Rate (%)	LP Fraction	Rate (%)	LP Fraction	Rate (%)
SMaLL100	5/10	2.160$_{0.02}$	6/13	0.054$_{0.00}$	1/2	0.025$_{0.02}$
M2M (S)	10/10	12.61$_{1.79}$	12/13	0.467$_{0.05}$	1/2	0.075$_{0.07}$
M2M (M)	7/10	12.22$_{2.41}$	7/13	0.172$_{0.05}$	0/2	0.000$_{0.00}$
M2M (L)	6/10	6.580$_{2.02}$	4/13	0.077$_{0.00}$	0/2	0.000$_{0.00}$

Table 4: Fraction of LPs on the non-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.

the corpora that were used to create the training data, we found reference translations that exactly match the toxic hallucinations. Additionally, we found that these hallucinations can propagate through model distillation, as evidenced by SMaLL100 generating copies of its teacher model's toxic hallucinations. This highlights the necessity of rigorously filtering training data to ensure safe use of these models.

5.3 Beyond English-Centric Translation

We shift our focus to translation directions that do not involve English, typically corresponding to directions with less supervision during training.

Trends are largely similar to English-centric directions. Table 4 reveals trends that largely mirror those observed in the English-centric setup: (i) hallucinations are more frequent in low-resource directions;[20] (ii) SMaLL100 significantly outperforms the M2M models in low-resource language pairs; and (iii) scaling up to M2M (L) consistently yields substantial improvements over the smaller M2M models in low- and mid-resource directions. Additionally, the trends related to hallucination types are also similar: Detached hallucinations are more prevalent in low-resource directions, while oscillatory hallucinations overwhelmingly dominate in mid- and high-resource ones.

Less supervised language pairs exhibit extremely high hallucination rates. As expected, models struggle more with hallucinations for directions with less or even no supervision during training, such as ro-hy and af-zu. For instance, M2M (M) hallucinates for nearly half of the translations in these directions.

Resource	SMaLL100	M2M (S)	M2M (M)	M2M (L)
Low	−0.019	−1.516	−1.317	−0.412
Mid	0.021	0.080	0.095	−0.013
High	−0.008	0.007	0.000	0.000

Table 5: Delta average hallucination rate at each resource level for Flores and TICO medical data. Positive values indicate higher rates for TICO.

5.4 Translation on Specialized Domains

We now study hallucinations in medical data using the TICO dataset. We compare hallucination rates with the Flores dataset in 18 directions.

Hallucinations are not exacerbated under medical domain data. Table 5 reveals that hallucination rates for the TICO data are not consistently higher than those observed for the Flores data. This finding diverges from previous work that investigated hallucinations in specialized domain data (Wang and Sennrich, 2020; Müller et al., 2020). We hypothesize that, unlike the smaller models typically trained on limited datasets from a single domain used in those works, the concept of ''domain shift'' is not as pronounced for M2M models. These models are not only much larger but, crucially, they are trained on a massive dataset containing over 7B sentences from various domains. This vast training set potentially mitigates the impact of domain shift and, consequently, reduces its influence on hallucinations.

6 Mitigation of Hallucinations through Fallback Systems

We now explore the potential of mitigating hallucinations and improving overall translation quality by employing a simple hybrid setup that can take advantage of multiple systems with possible complementary strengths. Put simply, we leverage an

[20] We considered the resource level of the language pair to be the smallest resource level between the two languages.

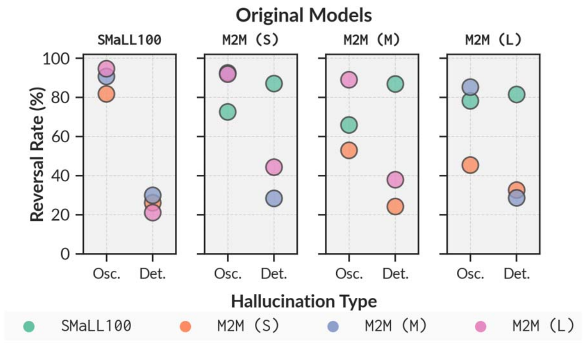


Figure 4: Reversal rates for oscillatory (Osc.) and detached (Det.) hallucinations when same-family systems (dots) are used as fallback to an original model (labeled above subplots).

alternative system as a fallback when the original model produces hallucinations.[21] Our analysis is focused on the broader English-centric evaluation setup studied in Section 5.2.

6.1 Employing Models of the Same Family as Fallback Systems

We begin by analyzing the performance of same-family models when employed as fallback systems for one another (e.g., using SMaLL100, M2M (M), and M2M (L) as fallbacks for M2M (S)).[22]

Detached hallucinations are particularly *sticky* across M2M models. Figure 4 reveals that when employing M2M models as fallback systems, reversal rates—percentage of hallucinations from the original system that are reversed by the fallback system—are consistently higher for oscillatory hallucinations than for detached hallucinations. These findings not only align with those in Guerreiro et al. (2023b), where oscillatory hallucinations were found to be less related to model defects, but also further highlight the close connection between detached hallucinations and the training data. This connection can help explain their *stickiness*: Since the M2M models share the same training data, reversing these hallucinations using other M2M variants as fallbacks is more challenging. Interestingly, we also observe that M2M (L) particularly struggles to reverse the detached hallucinations generated by its distilled counterpart SMaLL100, suggesting that model defects can persist and be shared during distillation.

Scaling up within the model family is not an effective strategy for mitigating hallucinations. In line with our findings in Section 5.2, Figure 4 shows that reversal rates using SMaLL100 as a fallback system are higher for detached hallucinations than for oscillatory hallucinations. Although SMaLL100 is a distilled M2M model, its training data, training procedure, and architecture differ from those of the M2M models. This distinction may make it a more complementary fallback system to other M2M models than simply scaling up within the same model family. This suggests that merely increasing the scale of models within the same family is not an effective strategy for mitigating hallucinations, and exploring alternative models with different training procedure and data may yield further improvements. We will analyze this alternative strategy in the next section.

6.2 Employing Diverse Fallback Systems

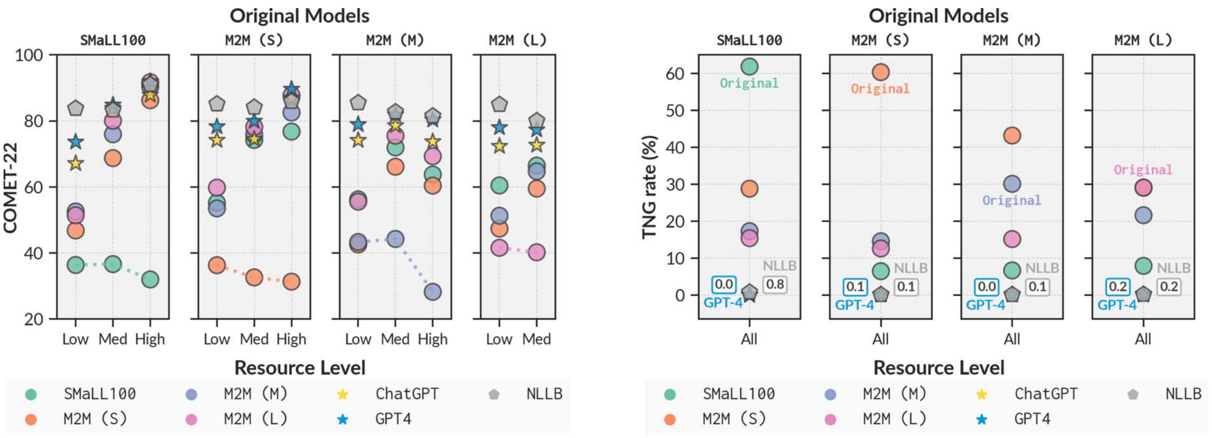
Building on the findings from the previous section, we will investigate how models outside the M2M family can be employed to mitigate hallucinations and improve translation quality. We will test this approach with two different models: (i) we will prompt the GPT models as detailed in Section 3,[23] and (ii) we will use a high-quality 3.3B parameter model from the NLLB family of NMT models (NLLB) (NLLB Team et al., 2022).

Diverse fallback systems can significantly improve translation quality. Figure 5a shows that diverse fallback systems can significantly enhance translation quality of originally hallucinated translations compared to same-family models. This improvement is most pronounced for low-resource directions, where both the LLMs and NLLB consistently boost translation quality. Moreover, NLLB generally outperforms the GPT models for low- and mid-resource language pairs, aligning with previous work that found that these models lag behind supervised models in these directions (Hendy et al., 2023). Nonetheless, even for these language pairs, GPT models exhibit superior performance to the dedicated M2M translation

[21] For consistency, we use the same detection approach described in Section 5.1.

[22] For simplicity, we consider the distilled SMaLL100 as a model from the M2M family.

[23] We remark again that GPT models may have been exposed to the evaluation data.



(a) Translation quality.

(b) Prevalence of oscillatory hallucinations.

Figure 5: We analyze overall translation quality improvements on the original model hallucinated translations (represented with dashed lines) across different resource levels via COMET-22 scores in (a), and overall prevalence of oscillatory hallucinations among the fallback translations in (b).

systems, further underscoring the limitations of relying on same-family models as fallbacks.

Oscillatory hallucinations are practically non-existent when employing diverse fallbacks. Figure 5b demonstrates another benefit of using diverse fallback systems: oscillatory hallucinations are almost completely eliminated. Consistent with our findings in Section 4, we observe that LLMs produce very few, if any, oscillations, slightly improving the rates obtained with NLLB. This provides further evidence that hallucinations obtained with LLMs exhibit different properties and surface forms. Investigating these differences is a relevant research direction for future work.

7 Conclusion

We have comprehensively investigated hallucinations in massively multilingual translation models, exploring a wide range of translation scenarios that remained overlooked in previous work.

Our analysis provided several key insights on the prevalence and properties of hallucinations across models of different scales and architectures, translation directions, and data conditions, including: the prevalence of hallucinations in low-resource languages and when translating out of English; the emergence of toxicity in hallucinations, which can be directly traced back to the training data; how model distillation may bring reduced hallucination rates compared to larger models; how scaling up within the same model family generally decreases the rate of hallucinations; and how LLMs produce qualitatively different hallucinations compared to conventional NMT models. Finally, we also examined how fallback systems can be employed to mitigate hallucinations. We found that hallucinations can be *sticky* and difficult to reverse when using models of the same family. However, diverse fallbacks with different training procedure and data can significantly improve translation quality and virtually eliminate pathologies such as oscillatory hallucinations.

Acknowledgments

We thank our action editor and the anonymous reviewers for their detailed and helpful feedback on this paper. We would like to thank Meta AI for open-sourcing the M2M models and maintaining libraries such as `stopes` (Andrews et al., 2022) and `nllb` (NLLB Team et al., 2022). The work is partially supported by the European Research Council (ERC StG DeepSPIN 758969), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (10039436 – UTTER), by the FCT through contract UIDB/50008/2020, and by the projects MAIA and NextGenAI (LISBOA-01-0247-FEDER-045909 and 2022-C05i0102-02). This work made use of HPC resources from GENCI-IDRIS (Grant 2022-AD01101838).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N19-1388`

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics. `https://aclanthology.org/2021.wmt-1.1`

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: The translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.nlpcovid19-2.5`

Pierre Andrews, Guillaume Wenzek, Kevin Heffernan, Onur Çelebi, Anna Sun, Ammar Kamran, Yingzhe Guo, Alexandre Mourachko, Holger Schwenk, and Angela Fan. 2022. stopes—modular machine translation pipelines. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 258–265, Abu Dhabi, UAE. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-demos.26`

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. `https://doi.org/10.48550/ARXIV.1907.05019`

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. `https://doi.org/10.48550/ARXIV.2205.03983`

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: The case of bloom. `https://doi.org/10.48550/ARXIV.2303.01911`

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Lan guage Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.acl-long.224`

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational

Linguistics. https://doi.org/10.18653/v1/W19-5361

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. http://arxiv.org/abs/2304.11158

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. https://doi.org/10.48550/ARXIV.2005.14165

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.236

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2202.07646

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. https://doi.org/10.48550/ARXIV.2204.02311

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. http://arxiv.org/abs/2205.12446

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.3

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1045

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1213

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. https://doi.org/10.48550/ARXIV.2010.11125

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. https://doi.org/10.18653/v1/2022.emnlp-main.599

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. https://doi.org/10.48550/ARXIV.2302.04166

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538. https://doi.org/10.1162/tacl_a_00474

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023a. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.770

Nuno M. Guerreiro, Elena Voita, and André 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, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.75

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? A comprehensive evaluation. https://doi.org/10.48550/ARXIV.2302.09210

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. https://doi.org/10.48550/ARXIV.2202.03629

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5506

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.1

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. https://doi.org/10.48550/ARXIV.2302.14520

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. In *Hallucinations in neural machine translation*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.616

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. https://aclanthology.org/2022.emnlp-main.616

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.571

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-research.14

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. https://doi.org/10.48550/ARXIV.2207.04672

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. https://doi.org/10.48550/ARXIV.2203.02155

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *ResearchGate preprint*. https://doi.org/10.18653/v1/2022.emnlp-main.225

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*. https://doi.org/10.18653/v1/2022.emnlp-main.225

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. http://arxiv.org/abs/2212.04356

Alec Radford and Karthik Narasimhan. 2018. In *Improving language understanding by generative pre-training*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.naacl-main.92`

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection. `https://doi.org/10.18653/v1/2022.findings-emnlp.379`

Ricardo Rei, José G. C. De Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. `https://aclanthology.org/2022.wmt-1.52`

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. `https://doi.org/10.18653/v1/P16-1009`

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P16-1009`

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D19-1331`

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.acl-long.859`

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.326`

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online. Association for Computational Linguistics. `https://aclanthology.org/2021.wmt-1.2`

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. `https://doi.org/10.48550/ARXIV.2301.07779`

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 1628–1639, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.148

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. https://doi.org/10.48550/ARXIV.2205.01068

A Appendix

A.1 Human Validation of Detection of Hallucinations under Perturbation

Detection of hallucinations under perturbation with the method that we adopted in our work (see Section 4.1) has been consistently used in all previous research (Lee et al., 2018; Raunak et al., 2021; Xu et al., 2023). However, as it has not been appropriately validated in previous works, we ran a human evaluation study to assess its outputs.

We collected annotations for over 200 translations across 10 languages.[24] Table 6 shows that the vast majority of detected translations ($\sim$ 87%) were indeed confirmed as hallucinations, containing content that is detached from the source text. Interestingly, this number is very aligned with the percentage of hallucinations under perturbation that would be detected with our detection approach for natural hallucinations ($\sim$ 85%; see Section 5.1). Moreover, we found that 9% of the flagged translations contained other errors, such as mistranslations, or more severe issues (e.g., undergeneration (Stahlberg and Byrne, 2019)). These findings validate the adopted detection method for hallucinations under perturbation in Section 4.

Correct	Incorrect	Hallucination	All
10 (4%)	19 (9%)	**194 (87%)**	223

Table 6: Human annotations on translations flagged by our detection method. Hallucinations are flagged by annotators when the translation contains text that is detached from the source.

A.2 Extractability of FLORES Samples with GPT Models

We want to analyze the extent to which GPT models may be able to extract the samples from the FLORES dataset used in Section 4. We use the definition of extractability proposed in Carlini et al. (2023) and later studied in Chowdhery et al. (2022) and Biderman et al. (2023). Under this definition, a string s is extractable with k tokens of context from a model f if there exists a (length-k) string p, such that the concatenation $[p||s]$ is contained in the training data for f, and f produces s when prompted with p using greedy decoding.[25] We follow the setup of Carlini et al. (2023), and set k to $L - 10$, where L is the length of the test sequence (with $L \geq 10$).[26]

We use XGLM (Lin et al., 2022), a 7.5B LLM trained on data predating the creation of FLORES as a baseline model that has not seen the translations in the target languages.[27]

Table 7 shows that GPT models can *only* extract a maximum of 2 ($\sim$ 1 in 1000) samples in the target languages, with most languages yielding zero extracted sentences. Rates are higher for English (source) sentences, which is to be expected: The English sentences from the FLORES dataset originate from Wikipedia, a regular source of training data for language models, whereas the target sentences were created specifically for the benchmark. Importantly, when compared to XGLM, the GPT models could only extract, at

[24] We hired annotators for each of the 10 languages (bn, el, tr, hi, tl, sw, ast, vi, he from FLORES; and ha from WMT). Overall, we annotated 223 samples, which amounts to more than half of all the detected translations. We adapted the guidelines for human annotation of hallucinations used in Dale et al. (2023), and will make them publicly available with all other resources. Annotators were sourced from Upwork, receiving compensation between \$20 and \$30 per hour.

[25] In the studies of Carlini et al. (2023), Chowdhery et al. (2022), and Biderman et al. (2023), the authors have access to the training data. We will assume that the samples from the FLORES dataset are included in the training data of GPT models.

[26] In Carlini et al. (2023), the model is required to emit training sequences, often documents, by generating significantly more tokens. For example, they set k to $L - 50$.

[27] This is important, as some test examples may be extracted, not because they have been memorized, but because they may be very similar to other training set examples (Chowdhery et al., 2022). We selected target languages from different resource levels that are supported by XGLM.

Model	Source	Target					
	en	sw	vi	id	nl	es	pt
XGLM	1 (0.05%)	0	1 (0.05%)	0	0	1 (0.05%)	0
ChatGPT	6 (0.30%)	0	2 (0.10%)	0	0	2 (0.10%)	0
GPT-4	9 (0.45%)	0	2 (0.10%)	0	0	2 (0.10%)	1 (0.05%)

Table 7: Counts and rate, in percentage, of *extractable* samples from the Flores dataset.

Model	Low		Mid					High
	ha†	hr	cs	is†	ja	uk	zh	ru
SMaLL100	0.000	0.000	0.000	0.000	0.087	0.000	0.000	0.000
M2M (S)	6.878	0.223	0.000	0.505	0.175	0.087	0.087	0.087
M2M (M)	1.058	0.111	0.000	0.000	1.400	0.000	0.000	0.087
M2M (L)	0.529	0.000	0.000	0.000	0.175	0.175	0.000	0.000
ChatGPT	1.587	0.334	0.000	0.168	0.000	0.087	0.000	0.262
GPT-4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 8: Average hallucination rate, in percentage, for each language pair on the WMT 2022 benchmarks.

the most, one additional sentence for some target languages.

Notably, even if minor train-test overlap exists, their influence on BLEU scoring differences on translations obtained with LLMs has been found to be small, or almost negligible, when translating out of English (Vilar et al., 2023; Chowdhery et al., 2022). Alongside the results in Table 7, we are confident that our analysis with the GPT models in Section 4 is valuable and valid, irrespective of any hypothetical, albeit small, train-test overlap. Nevertheless, in the next section, we experiment on benchmarks created after the training data cutoff of these LLMs.[28] Crucially, the trends in the results closely follow those reported in Section 4.

A.3 Hallucinations under Perturbation on WMT 2022 Benchmarks

We now investigate hallucinations under perturbation for WMT 2022 benchmarks (Kocmi et al., 2022) created after the cutoff data of the training data of GPT models. Additionally, to include more languages to our analysis, we also experiment with WMT 2021 benchmarks (Akhbardeh et al., 2021) (for Hausa, ha, and Icelandic, is), as they are less likely to overlap with the training data of GPT models. We apply the same perturbations and detection method detailed in Section 4.1.[29]

Trends with GPT models largely mirror those reported on the Flores dataset. Table 8 reveals trends that align closely with those from Section 4: (i) SMaLL100 consistently exhibits lower hallucination rates compared to other NMT models, including its teacher model, M2M (L); (ii) GPT-4 shows impressive performance across the board, with no instances of hallucinations across all language pairs; and (iii) ChatGPT may hallucinate more than NMT models for some language pairs, generating hallucinations that are qualitatively different from those produced by these models, often consisting of failed attempts to translate (e.g., *''Sorry, but my language module cannot translate English text into Russian.''*).

[28] This chronology suggests that the train-test overlap risk is reduced. However, as the training data is not available, we cannot entirely rule out overlap with the test data.

[29] We do not introduce perturbations coming from a speech recognition system, as audio recordings of these test sets are not available. We also advise against comparing absolute hallucination rates between two language pairs, as the test sets—and thus, the set of hallucination candidates—for each language pair have distinct sizes.