Can LLMs Detect Intrinsic Hallucinations in Paraphrasing and Machine Translation?

Evangelia Gogoulou[1,3] **Shorouq Zahra**[1,2] **Liane Guillou**[4,5] **Luise Dürlich**[1,2] **Joakim Nivre**[1,2]

[1]RISE Research Institutes of Sweden [2]Uppsala University
[3]KTH Royal Institute of Technology [4]University of Edinburgh [5]Aveni

Abstract

A frequently observed problem with LLMs is their tendency to generate output that is nonsensical, illogical, or factually incorrect, often referred to broadly as "hallucination". Building on the recently proposed HalluciGen task for hallucination detection and generation, we evaluate a suite of open-access LLMs on their ability to detect intrinsic hallucinations in two conditional generation tasks: translation and paraphrasing. We study how model performance varies across tasks and languages and we investigate the impact of model size, instruction tuning, and prompt choice. We find that performance varies across models but is consistent across prompts. Finally, we find that NLI models perform comparably well, suggesting that LLM-based detectors are not the only viable option for this specific task.

1 Introduction

The introduction of large language models (LLMs) has revolutionised the field of natural language processing (NLP). State-of-the-art LLMs have demonstrated excellent language generation capabilities. in conversational AI (Zhao et al., 2024), as well as strong performance on more specific NLP tasks like summarisation (Pu et al., 2023), open-domain question answering (Kamalloo et al., 2023), sentiment analysis (Zhang et al., 2024), and machine translation (Kocmi et al., 2023). Despite this success, LLMs are prone to producing output that is fluent and grammatical, but semantically inadequate or factually incorrect, a phenomenon broadly referred to within the NLP community as "hallucination". The impact of hallucinations by LLMs may be severe in downstream applications where accurate output is mission critical, or where hallucination leads to erroneous decisions with negative consequences that directly impact humans e.g. in the medical or legal domain. In many cases, it may be infeasible to have a human in the loop, or it may be difficult for humans to identify hallucinations, which motivates the need for automated methods for detection and evaluation.

In this paper, we aim to discover whether LLMs can be used to detect hallucinated content, focusing on a special case of what Ji et al. (2023) call intrinsic hallucinations, that is, cases where the output is deficient with respect to a particular input and where the deficiency can be detected given only the input and output.[1] More precisely, for the tasks of paraphrasing and machine translation, we define a hallucination to be an output, or hypothesis, that is not entailed by the input, or source.

We build upon our previous work from the ELOQUENT Lab at CLEF 2024 (Dürlich et al., 2024), specifically the HalluciGen task, where we asked participants to apply LLMs to the task of detecting and generating hallucinations. We extend the work from the shared task with a series of experiments in prompting open-access LLMs to detect hallucinations, framing it as a contrastive challenge task: given a source sentence, and a pair of hypotheses, the model should detect which one contains a hallucination. We evaluate the LLMs on hallucination detection in paraphrase generation and translation, as defined in the HalluciGen task (Dürlich et al., 2024).

Through a systematic investigation of model performance on the hallucination detection task, we address the following questions:

- How does model performance differ across target languages?
- Does increased model parameter size improve performance?
- Does instruction tuning improve performance?
- Does the language and formulation of the prompt matter?

[1]This in contrast to extrinsic hallucinations, where additional information such as world knowledge is required to detect the deficiency.

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 161–177

2 Background and Related Work

Two concepts that are often used to characterise different types of hallucinations are *faithfulness* and *factuality*. Faithfulness means being consistent with a given source or input and has long been used as an evaluation criterion in conditional generation tasks like machine translation; a faithfulness hallucination is therefore any output that lacks such consistency, regardless of whether it is factually correct. By contrast, factuality means corresponding to real-world knowledge, and a factuality hallucination is therefore any output that makes a false claim, regardless of context and input. A related distinction is made between *intrinsic* and *extrinsic* hallucinations, where the former can be detected from the input and output of a system alone, while the latter requires more information (Ji et al., 2023).

Prior work has mostly focused on building systems to detect factuality hallucinations. For example, Li et al. (2023) introduce a benchmark targeting cases of factual hallucinations in the context of question-answering, knowledge-grounded dialogue, and summarisation. Aside from the HalluciGen task, the closest work to ours is the SHROOM shared task (Mickus et al., 2024) from SEMEVAL 2024. SHROOM defines hallucinations as cases when the hypothesis cannot be inferred from its semantic reference. Despite the similarity with our definition, there is a significant difference in how the hallucinations are constructed. In SHROOM they are generated by models prompted to solve the specific task scenario, whilst we mostly construct hallucinations manually based on specific categories of errors; by switching gender, negation, or tense, replacing words with their antonyms, by substituting named entities, numbers, dates, and currencies, and by making superfluous additions. The two tasks also differ in terms of their coverage of NLP tasks and target languages. SHROOM includes the additional task of definition modeling; HalluciGen covers an extra language for paraphrase but has limited coverage for machine translation.

There is limited evidence so far on the effectiveness of using LLMs for detecting hallucinations. Li et al. (2023) find that LLMs, including Llama2 and ChatGPT, perform poorly on the task of identifying hallucinations that have been generated by LLMs to be factually incorrect, in English question-answering and summarisation. According to the HalluciGen task results (Dürlich et al., 2024), GPT-4 and LLM majority voting approaches outperform smaller English-centric models such as Llama3-8b and Gemma-7b. Similar conclusions emerge from SHROOM, where submissions based on GPT-4 or model ensembling exhibit the strongest performance. Model fine-tuning on SHROOM training data is another successful approach.

Conversely, textual entailment classifiers have been utilised for detecting faithfulness hallucinations. Maynez et al. (2020) argue that textual entailment classifiers correlate with the faithfulness of summarised texts, making NLI models a suitable candidate for automatic evaluation. Textual entailment has also been applied to the evaluation of translations. Padó et al. (2009) address the issue of robustness in MT evaluation and propose a metric based on features motivated by textual entailment for "assessing the meaning equivalence between reference and hypothesis". Similarly, Marouani et al. (2020) developed a metric directly incorporating a textual entailment system, where a perfect translation pair would score highly in entailment in both directions (noting that omissions and additions can adversely affect entailment).

Manakul et al. (2023) compare the performance of both approaches by introducing SelfCheckGPT, which detects sentence-level hallucinations using generative LLM prompting, LLM probabilities, and NLI models. Interestingly, their experimental results show that LLM prompting outperforms the NLI-based method only by a small margin, and both outperform all other SelfCheckGPT methods and baselines. Likewise, Kryscinski et al. (2020) demonstrate that classifiers trained on MNLI (Williams et al., 2018) can perform well on factuality hallucination detection tasks. However, they are outperformed by similar classification models trained on a set of synthetically generated hallucinations (through sentence negation, entity swapping, and noise insertion), with the objective of classifying a source document and claim sentence as either "consistent" or "inconsistent". Additionally, NLI-based methods yield promising results for high-resource languages in multilingual setups, often outperforming other lexical metrics (like ROUGE), especially for intrinsic hallucinations where the hypothesis would clearly contradict the source (Kang et al., 2024).

The ability of NLI models to detect intrinsic hallucinations is arguably unsurprising as they must "handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lexical and syntactic ambiguity" (Williams et al.,

2018) to successfully predict entailment, contradiction, and neutral relations between sentence pairs.

3 Dataset Description

The HalluciGen detection task (Dürlich et al., 2024) covers the two following scenarios:

- **Paraphrase Generation**: The model is presented with two possible paraphrases of a given source sentence in English (en) and Swedish (sv).
- **Machine Translation**: Given a sentence in a source language, the model is presented with two possible translations in the target language; English-German (en⇔de) and English-French (en⇔fr), in both translation directions.

Each example in the dataset consists of a source sentence (src), a good hypothesis ($hyp+$), and an incorrect hypothesis containing an intrinsic hallucination ($hyp-$). The criterion for a hypothesis to contain such a hallucination is that it is not entailed by the source sentence, which in turn means that it must contain some additional or contradictory information with respect to the source. This may be due to additions, substitutions, negations, or other phenomena that break the inference relation. Note that this definition is a relaxation of the definition in Ji et al. (2023), where intrinsic hallucinations are required to explicitly contradict the source. Note also that a hypothesis that does not entail the source sentence is not considered a hallucination, despite being an imperfect paraphrase/translation, as long as it is still entailed by the source. For example, if the source is "it is cold and wet", then "it is cold and windy" and "it is not cold and wet" are both considered hallucinations, but "it is cold" is not.

Each hallucinated hypothesis belongs to one of eleven categories, defined by the type of error or addition that breaks the entailment relation: addition, named-entity, number, conversion, date, gender, pronoun, antonym, tense, negation, natural. The last category refers to hallucinated responses by LLMs that did not fit into any of the other above categories. Examples of each hallucination category for the paraphrase task can be found in Table 4 in Appendix A, and the frequency statistics of the hallucination categories in Appendix B. All datasets are available on Huggingface.[2]. The dataset creation process for the translation and paraphrase scenarios is summarised below and described in full in Dürlich et al. (2024).

3.1 Paraphrase Generation

The English dataset consists of 138 examples from the SHROOM training set for the paraphrase generation subtask (Mickus et al., 2024). For the Swedish dataset, 139 examples from the SweParaphrase test data were used (Berdicevskis et al., 2023), consisting of sentence pairs together with their degree of semantic similarity, and the Swedish part of the Finnish Paraphrase Corpus (Kanerva et al., 2021), which consists of paraphrase hypothesis pairs and a label indicating the degree of paraphrase relation. The selected examples have the highest similarity (SweParaphrase), or are paraphrase equivalents (Finnish Paraphrase Corpus).

Mixtral-8x7B-instruct (Jiang et al., 2024) and GPT-SW3-6.7B-instruct (Ekgren et al., 2024) were used to automatically generate a paraphrase hypothesis for the first sentence of each pair, after which all examples were manually annotated in two steps. The annotators first determined whether the generated hypothesis is an intrinsic hallucination with respect to the source (see Appendix H). Then for those hypotheses not marked as hallucinations, the annotators manually constructed a hallucination based on one of the first ten categories (i.e. excluding natural hallucinations). The hypotheses marked as hallucinations were assigned to one type, or the natural type if they did not correspond to any specific hallucination phenomenon.

The test set for each language consists of 119 examples, with 16 additional trial examples for English and 20 for Swedish. We use Krippendorff's alpha to compute inter-annotator agreement on binary classification (hallucination or not) of the examples by three annotators. We observe high agreement: 0.90 for English, 0.88 for Swedish. The annotation guidelines are provided in Appendix H.

3.2 Machine Translation

Dürlich et al. (2024) leveraged ACES (Amrhein et al., 2022), a contrastive challenge set for evaluating machine translation metric performance on a range of translation accuracy errors. ACES examples consist of a source sentence, a pair of good/incorrect translation hypotheses, a reference translation, and a label denoting the error phenomenon in the incorrect translation. As ACES already contains examples for en⇔fr and en⇔de for most of the hallucination categories (except tense and negation) the majority of the HalluciGen

[2] https://huggingface.co/datasets/NLP-RISE/HalluciGen

dataset examples were sampled directly. For the tense and negation categories, new examples were constructed using the PAWS-X dataset (Yang et al., 2019) of adversarial paraphrases.

For each language direction, 100 test examples were sampled from the categories of ACES aiming for a uniform distribution across these categories as much as possible. Additionally, 10 trial examples were selected for each language direction.

4 Experimental Setup

4.1 Models

We evaluate a range of different model families, which differ in the type and amount of pre-training language data. From each family, we select multiple model variants that differ in model size and/or presence of instruction tuning. This enables the systematic study of those two factors in relation to the ability of the model to detect hallucinations. We select a number of variants from the **Llama3** (Dubey et al., 2024), **Mixtral** (Jiang et al., 2024), **EuroLLM**, and **GPT-SW3** (Ekgren et al., 2024) model families. The full list of models is found in Appendix E. The GPT-SW3 models are evaluated only in the paraphrase scenario, while the rest are used for both scenarios.

As our goal is to evaluate the inherent ability of the base model to detect hallucinations, we refrain from model fine-tuning on relevant data and few-shot prompting. After experimentation on the trial sets, the following generation parameters were used for all models: temperature = 0.1, top-k sampling = 20, maximum number of generated tokens = 5. Information about the computational efficiency of our experiments can be found in Appendix G.

4.2 Prompting

To investigate how model performance depends on the specific formulation of the prompt, we experiment with six different prompting strategies, exemplified in Table 1. The prompts differ with respect to whether they explicitly mention the term "hallucination" (Prompts 1–3 vs. 4–6) and whether they include an explicit definition of the concept of hallucination (Prompts 1–2 vs. 3–6). Prompts 4–6 (which contain neither the term "hallucination" nor an explicit definition) use formulations that to different degrees approximate the notion of hallucination with terms like "contradicts", "supports" and "bad". Note that the formulation with "support" inverts the task by prompting the model to identify the good hypothesis rather than the hallucination, which needs to be handled in post-processing to make sure that the evaluation is correct (see Appendix D). An additional variable is the language of the prompt: we experiment with prompting in English versus the language of the source sentence (which in the case of paraphrase is also the target language). Prompts in Swedish, French, and German can be found in Table 6 in Appendix C.

In addition to the base prompts, all models receive a near identical set of instructions to provide only "hyp1" or "hyp2" as acceptable answers and to start the text generation with "The answer is:" (or a similar phrase). Differences in the additional prompt instructions are minimal; they vary only by language or phrasing depending on the model. Though we did not prompt the models to do so, they sometimes provide explanations of the output.

4.3 Evaluation

All models are evaluated with respect to the gold labels in the datasets, using the F1 metric. The model output first undergoes simple rule-based post-processing to check for produced labels in a number of variations and map them to $hyp1$ or $hyp2$ (e.g. "hypothesis 1" or "första" for hypothesis 1, and "hypothesis 2" or "zweite" for hypothesis 2). Model outputs are considered invalid in cases where the model produces either no label at all or a label outside of the allowed set: $\{hyp1, hyp2\}$. Examples of outputs produced during the experiments can be found in Table 1. The post-processing is described in more detail in Appendix D.

4.4 NLI Baseline

As baselines, we use NLI models, which are computationally inexpensive and trained specifically for predicting textual entailment. NLI models typically classify a sentence pair into one of three classes: entailment, neutral, and contradiction. We selected two multilingual zero-shot NLI models with no "neutral" label, meaning they only predict the textual entailment between a premise and a hypothesis. The baseline used for all scenarios is BGE-M3-ZEROSHOT-V2.0, a multilingual zero-shot XLM-RoBERTa model based on BGE M3-Embeddings (Chen et al., 2024). An additional NLI baseline for the Swedish paraphrase scenario is SCANDI-NLI-LARGE (Nielsen, 2022), which is trained on Swedish, Danish, and Norwegian data. We first predict "entailment" and "not_entailment" class scores between the source sentence and each hypothesis.

Prompt Name	Prompt	Example output
Prompt 1	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which of the two is a hallucination of the src. Hallucination means that the hypothesis is not logically supported by the src.	"hypothesis1" ⇒ hyp1
Prompt 2	You are an AI judge specialised in <scenario> detection. Your task is the following: Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which of the two is a hallucination of the src. Hallucination means that the hypothesis is not logically supported by the src.	"The answer is hyp2" ⇒ hyp2
Prompt 3	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which of the two is a hallucination of the src.	"second" ⇒ hyp2
Prompt 4	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which one of the two logically contradicts the src.	"both" ⇒ invalid
Prompt 5	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which one of the two supports the src.	"2" ⇒ hyp2 ⇒ hyp1*
Prompt 6	Given a source sentence (src) and two paraphrase hypotheses (hyp1 and hyp2), judge which of the two is a bad <scenario> of the src.	"Hypothesis" ⇒ invalid
Prompt 6	You are an AI judge with expertise in machine translation. Given a source sentence (src) and two translation hypotheses (hyp1 and hyp2), your task is to judge which of the two is a bad translation of the source.	"It's hard to say" ⇒ invalid

Table 1: Prompt formulations in English tested on all models. For prompts 1-5 <scenario> is replaced with "paraphrase" or "translation". The last column shows example of generated outputs (translated to English when needed) and the label extracted by post-processing. These examples occur across all prompt variations and are not limited to the prompt they appear next to. *Note that Prompt 5 is a special case where the label is flipped.

We infer the label based on the predicted entailment value for each of the two hypotheses. More details can be found in Section F in the Appendix.

The default configurations are used for both models and each pair ($source$+$hyp1$ / $source$+$hyp2$). For the translations, the BGE-M3-ZEROSHOT-V2.0 NLI model receives two sentences in two different languages as input (one in English, and one in French or German) in both directions.

5 Results

Tables 2 and 3 present model scores for different prompt formulations and prompt languages in the paraphrase and translation scenarios. Overall, we observe that performance varies considerably between models. We also note that the NLI baseline is hard to beat, especially in the paraphrase scenario and for translation from French to English. This corroborates the findings of Dürlich et al. (2024).

5.1 Paraphrase

For English paraphrases, we observe that META-LLAMA-3-70B-INSTRUCT has the strongest overall performance, although with three of the prompts it does not beat the NLI baseline. The competitive performance of the NLI baseline is even more apparent in the Swedish paraphrase scenario, where the best-performing LLMs (META-LLAMA-3-70B-INSTRUCT and MIXTRAL-8X7B-INSTRUCT) are outperformed by the NLI baseline, irrespective of the prompt used. All GPT-SW3 models perform poorly for both Swedish and English. A striking observation is that the performance of GPT-SW3-20B-INSTRUCT reaches the low F1 score of 0.07 for Prompt 2 for Swedish. When prompted with "You are an AI judge specialised in …", GPT-SW3-20B-INSTRUCT provides mostly invalid answers. EUROLLM-1.7B-INSTRUCT exhibits comparable performance with the GPT-SW3 models on English paraphrase, and even surpasses them on Swedish paraphrase. The latter is surprising given the larger amount of Swedish data in the GPT-SW3 models. Lastly, the performance of EUROLLM-1.7B is generally on par with GPT-SW3-20B.

5.2 Machine Translation

In the Machine Translation scenario, we again observe stronger performance for MIXTRAL-8X7B-INSTRUCT and META-LLAMA-3-70B-INSTRUCT compared with EUROLLM-1.7B-INSTRUCT. In contrast with the paraphrase scenario, where we observe that the NLI baseline often outperforms even the strongest LLMs, for translation we almost see the opposite: the NLI baseline is outperformed by either META-LLAMA-3-70B-INSTRUCT or MIXTRAL-8X7B-INSTRUCT for every language direction except fr⇒en. One obvious difference is that whilst the paraphrase task is monolingual, the cross-lingual nature of the translation task adds

English paraphrase								
BGE-M3-ZEROSHOT-V2.0	**0.90**							
LLM	**PLg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Avg ± SD**
META-LLAMA-3-8B-INSTRUCT	en	0.43	0.44	0.35	0.37	0.87	0.60	0.51 ± 0.20
META-LLAMA-3-70B-INSTRUCT	en	**0.84**	**0.92**	0.69	**0.88**	**0.94**	**0.91**	**0.86** ± 0.09
META-LLAMA-3-70B	en	0.70	0.58	0.59	0.70	0.63	0.81	0.67 ± 0.09
MIXTRAL-8X7B-INSTRUCT	en	0.76	0.79	**0.81**	0.80	0.82	0.86	0.81 ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.48	0.77	0.50	0.41	0.85	0.76	0.63 ± 0.19
EUROLLM-1.7B-INSTRUCT	en	0.32	0.41	0.28	0.33	0.57	0.29	0.37 ± 0.11
EUROLLM-1.7B	en	0.45	0.45	0.46	0.45	0.22	0.45	0.41 ± 0.09
GPT-SW3-20B-INSTRUCT	en	0.45	0.07	0.45	0.44	0.22	0.44	0.35 ± 0.16
GPT-SW3-20B	en	0.55	0.44	0.48	0.50	0.31	0.52	0.47 ± 0.09
GPT-SW3-40B	en	0.27	0.22	0.31	0.22	0.50	0.23	0.29 ± 0.11
Swedish paraphrase								
BGE-M3-ZEROSHOT-V2.0	**0.92**							
SCANDI-NLI-LARGE	**0.92**							
LLM	**PLg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Avg ± SD**
META-LLAMA-3-8B-INSTRUCT	en	0.49	0.56	0.49	0.53	0.58	0.50	0.52 ± 0.04
	sv	0.40	0.47	0.45	0.42	0.69	0.49	0.49 ± 0.10
META-LLAMA-3-70B-INSTRUCT	en	0.72	**0.86**	0.62	0.76	0.80	0.78	0.76 ± 0.04
	sv	**0.79**	0.81	0.46	0.65	**0.83**	0.83	0.73 ± 0.03
META-LLAMA-3-70B	en	0.54	0.45	0.55	0.63	0.56	0.63	0.56 ± 0.07
	sv	0.36	0.32	0.33	0.41	0.57	0.50	0.42 ± 0.10
MIXTRAL-8X7B-INSTRUCT	en	**0.79**	0.84	**0.85**	0.80	0.81	**0.86**	**0.83** ± 0.05
	sv	0.78	0.75	0.74	**0.88**	0.79	0.66	0.77 ± 0.08
MIXTRAL-8X22B-INSTRUCT	en	0.44	0.71	0.46	0.39	0.77	0.69	0.58 ± 0.17
	sv	0.38	0.34	0.28	0.40	0.79	0.09	0.38 ± 0.23
EUROLLM-1.7B-INSTRUCT	en	0.62	0.62	0.55	0.63	0.39	0.60	0.57 ± 0.01
	sv	0.34	0.33	0.33	0.33	0.32	0.33	0.33 ± 0.01
EUROLLM-1.7B	en	0.34	0.32	0.34	0.34	0.33	0.34	0.34 ± 0.00
	sv	0.33	0.34	0.33	0.34	0.33	0.33	0.33 ± 0.00
GPT-SW3-20B-INSTRUCT	en	0.33	0.14	0.33	0.33	0.32	0.33	0.30 ± 0.08
	sv	0.01	0.04	0.03	0.04	0.32	0.33	0.13 ± 0.15
GPT-SW3-20B	en	0.33	0.15	0.33	0.40	0.33	0.32	0.31 ± 0.08
	sv	0.39	0.33	0.37	0.35	0.32	0.36	0.35 ± 0.03
GPT-SW3-40B	en	0.43	0.34	0.5	0.41	0.45	0.52	0.44 ± 0.06
	sv	0.45	0.39	0.53	0.50	0.41	0.40	0.45 ± 0.06

Table 2: Test set results for the paraphrase scenario in English and Swedish: F1 scores. Baseline models have a single score. For all other models, we report scores for different combinations of prompt language (PLg) and prompt formulation (P1–P6), as well as (Avg) and standard deviation (SD). Boldface marks highest score per column.

complexity, as the model not only needs to perform the NLI task but also implicit translation. As translation examples are likely present in pre-training data, and possibly addressed by subsequent instruction-tuning, this may give LLMs an edge over NLI models. Further investigation is needed to determine whether this is the case.

6 Discussion

The results presented in Section 5 support the use of LLMs, and also NLI models, for the hallucination detection task. We now discuss the differences in performance across target languages as well as the effects of model size, instruction tuning, and the language and formulation of the prompts.

6.1 Research Questions

How does model performance on hallucination detection differ between target languages? We find that the capability of the model to detect hallucinations is generally consistent between target languages, with often a slight performance benefit for English source sentences. This is not surprising given that English is most likely the dominant language in the data used for pre-training and instruction tuning of the models. Two exceptions are GPT-SW3-40B and EUROLLM-1.7B-INSTRUCT. Both have better performance on Swedish than English, despite being trained on larger amounts of English data compared to Swedish. In addition, it is observed that EUROLLM-1.7B-INSTRUCT outperforms all three GPT-SW3 models on the Swedish paraphrase scenario, despite the limited amount of Swedish pre-training data in the former model. This indicates that the amount of target language data used in pre-training is not the sole factor contributing to the model performance on hallucination detection in languages other than English.

Does increased model parameter size lead to better performance? We compare the performance of models with different numbers of parameters belonging to the same family. For Llama3 we observe that model size has a clear impact, with the

Translation en⇒fr								
BGE-M3-ZEROSHOT-V2.0	0.82							
LLM	**PLg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Avg ± SD**
META-LLAMA-3-8B-INSTRUCT	en	0.74	0.77	0.66	0.71	0.83	0.73	0.74 ± 0.06
META-LLAMA-3-70B-INSTRUCT	en	**0.85**	**0.89**	0.81	**0.88**	**0.86**	**0.90**	**0.87** ± 0.03
META-LLAMA-3-70B	en	0.69	0.73	0.70	0.74	0.49	0.74	0.68 ± 0.10
MIXTRAL-8X7B-INSTRUCT	en	0.81	0.86	**0.85**	0.78	0.83	0.80	0.82 ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.41	0.68	0.57	0.44	0.74	0.45	0.56 ± 0.15
EUROLLM-1.7B-INSTRUCT	en	0.34	0.44	0.49	0.40	0.60	0.49	0.46 ± 0.09
EUROLLM-1.7B	en	0.44	0.42	0.44	0.43	0.23	0.43	0.40 ± 0.08
Translation fr⇒en								
BGE-M3-ZEROSHOT-V2.0	**0.88**							
LLM	**PLg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Avg ± SD**
META-LLAMA-3-8B-INSTRUCT	en	0.62	0.63	0.53	0.60	0.73	0.57	0.61 ± 0.07
	fr	0.33	0.40	0.30	0.43	0.80	0.73	0.50 ± 0.21
META-LLAMA-3-70B-INSTRUCT	en	0.67	0.80	0.53	**0.84**	**0.81**	0.78	0.74 ± 0.12
	fr	0.80	0.80	0.73	**0.84**	**0.81**	0.80	**0.80** ± 0.04
META-LLAMA-3-70B	en	0.63	0.70	0.58	0.68	0.61	0.66	0.64 ± 0.05
	fr	0.50	0.62	0.41	0.41	0.51	0.75	0.53 ± 0.13
MIXTRAL-8X7B-INSTRUCT	en	0.80	**0.82**	0.78	0.83	**0.81**	**0.81**	**0.80** ± 0.02
	fr	**0.81**	0.77	**0.85**	0.78	0.80	0.78	**0.80** ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.39	0.56	0.46	0.56	0.72	0.41	0.53 ± 0.15
	fr	0.07	0.26	0.05	0.13	0.53	0.34	0.24 ± 0.20
EUROLLM-1.7B-INSTRUCT	en	0.40	0.52	0.46	0.40	0.38	0.51	0.45 ± 0.06
	fr	0.35	0.36	0.32	0.34	0.31	0.35	0.34 ± 0.01
EUROLLM-1.7B	en	0.35	0.35	0.35	0.36	0.31	0.35	0.35 ± 0.02
	fr	0.35	0.34	0.35	0.34	0.31	0.34	0.34 ± 0.01
Translation en⇒de								
BGE-M3-ZEROSHOT-V2.0	0.73							
LLM	**PLg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Avg ± SD**
META-LLAMA-3-8B-INSTRUCT	en	0.56	0.62	0.48	0.57	0.79	0.60	0.60 ± 0.10
META-LLAMA-3-70B-INSTRUCT	en	0.69	**0.87**	0.68	**0.75**	0.83	**0.85**	0.78 ± 0.08
META-LLAMA-3-70B	en	0.65	0.70	0.61	0.65	0.54	0.81	0.66 ± 0.09
MIXTRAL-8X7B-INSTRUCT	en	**0.82**	0.79	**0.78**	**0.75**	0.84	0.79	**0.79** ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.49	0.75	0.64	0.57	**0.81**	0.59	0.65 ± 0.14
EUROLLM-1.7B-INSTRUCT	en	0.33	0.45	0.40	0.41	0.53	0.46	0.43 ± 0.07
EUROLLM-1.7B	en	0.42	0.41	0.42	0.42	0.24	0.42	0.39 ± 0.07
Translation de⇒en								
BGE-M3-ZEROSHOT-V2.0	0.78							
LLM	**PLg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Avg ± SD**
META-LLAMA-3-8B-INSTRUCT	en	0.56	0.58	0.46	0.52	0.79	0.47	0.57 ± 0.12
	de	0.41	0.36	0.19	0.48	0.80	0.67	0.49 ± 0.22
META-LLAMA-3-70B-INSTRUCT	en	0.66	0.85	0.60	0.82	0.81	**0.85**	0.77 ± 0.11
	de	0.53	**0.87**	0.20	**0.86**	**0.83**	0.83	0.69 ± 0.27
META-LLAMA-3-70B	en	0.56	0.57	0.50	0.55	0.67	0.60	0.58 ± 0.06
	de	0.34	0.72	0.30	0.38	0.67	0.56	0.49 ± 0.18
MIXTRAL-8X7B-INSTRUCT	en	0.75	0.82	**0.85**	0.85	0.81	0.84	**0.82** ± 0.04
	de	**0.81**	0.80	0.81	0.77	0.84	0.62	0.77 ± 0.08
MIXTRAL-8X22B-INSTRUCT	en	0.43	0.58	0.42	0.56	0.79	0.37	0.54 ± 0.18
	de	0.18	0.38	0.33	0.19	0.76	0.57	0.41 ± 0.24
EUROLLM-1.7B-INSTRUCT	en	0.22	0.23	0.21	0.22	0.46	0.21	0.26 ± 0.10
	de	0.20	0.22	0.22	0.22	0.45	0.22	0.26 ± 0.10
EUROLLM-1.7B	en	0.45	0.39	0.41	0.48	0.30	0.47	0.42 ± 0.07
	de	0.24	0.22	0.28	0.25	0.46	0.22	0.28 ± 0.09

Table 3: Test set results for the translation scenario in all language pairs: F1 scores. Baseline models have a single score. For all other models, we report scores for different combinations of prompt language (PLg) and prompt formulation (P1–P6), as well as (Avg) and standard deviation (SD). Boldface marks highest score per column.

larger META-LLAMA-3-70B-INSTRUCT model outperforming the smaller META-LLAMA-3-8B-INSTRUCT model, typically by a large margin. We see the same pattern for GPT-SW3, but only for Swedish, where GPT-SW3-40B consistently outperforms the smaller GPT-SW3-20B. The opposite trend is observed for the Mixtral models: increasing the model size from 8x7b to 8x22b consistently results in worse performance across all scenarios.

Does instruction tuning lead to better performance? In the case of the Llama3 family, we observe a clear performance improvement in using the instruction-tuned variant over the base META-LLAMA-3-70B in both scenarios and for all languages. The opposite is observed for GPT-SW3, with GPT-SW3-20B consistently outperforming the instruction-tuned variant on both paraphrase scenarios. This could be due to the absence of NLI examples in the instruction-tuning corpus used for

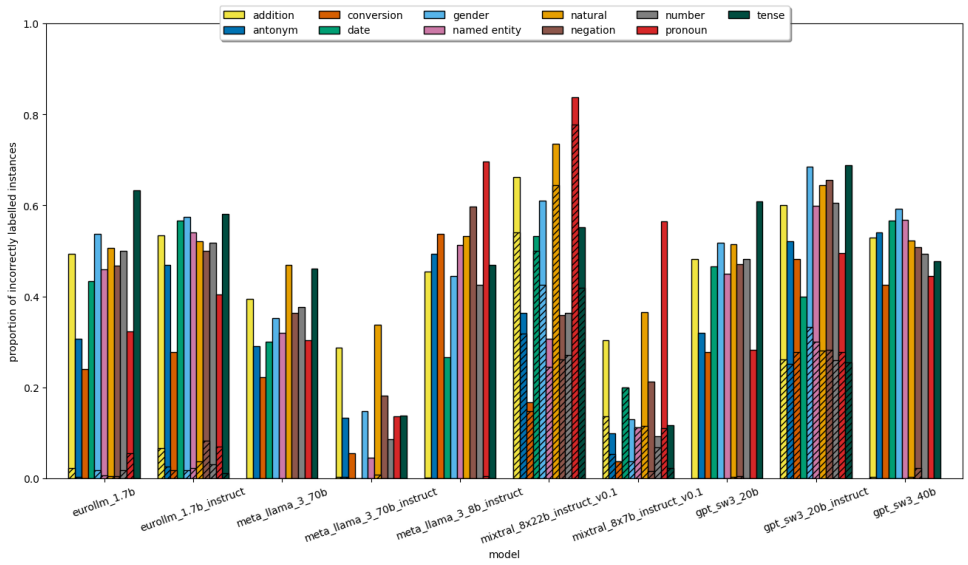

Figure 1: The average proportion of incorrectly labeled source-hyp **paraphrase pairs** (averaged over all prompts and prompt and data language combinations) filtered by hallucination category. Here, the hatch represents the proportion of outputs that were invalid (i.e. falling outside {$hyp1$, $hyp2$}).

GPT-SW3-20B-INSTRUCT (Ekgren et al., 2024). The instruction-tuned variant of EUROLLM-1.7B performs better for Swedish paraphrase and fr⇒en translation, while the reverse is true for English paraphrase and de⇒en translation. This may be attributed to the model's limited capacity, which restricts its ability to fully integrate the instruction tuning data. Overall, we do not find conclusive evidence that instruction tuning improves performance, as the results differ between model families, trained on different instruction tuning datasets.

Does the language and formulation of the prompt matter? We investigate the effect of non-English prompts for Swedish paraphrase and fr⇒en and de⇒en translation. As indicated by the difference in average model performance between prompt languages in Tables 2 and 3, the choice of prompt language matters, with English being overall the best-performing prompt language. This is not surprising given that all models under study have likely been trained on large amounts of English. One exception is Swedish paraphrase, where GPT-SW3-20B-INSTRUCT performs best with Swedish prompts. The same holds for META-LLAMA-3-70B-INSTRUCT, which performs best when prompted in French for fr⇒en translation.

We now investigate whether individual model performance varies with the prompt choice, considering the standard deviation values in Tables 2 and 3. Overall, performance remains stable across prompt variations, but certain cases stand out: MIXTRAL-8X22B-INSTRUCT is significantly unstable across all scenarios, with Prompt 5 (no mention of "hallucination" and use of "supports" instead of "contradicts") consistently performing best. The same partially holds for META-LLAMA-3-8B-INSTRUCT. Additionally, prompts mentioning "hallucination" (Prompts 1–3) tend to negatively impact performance for MIXTRAL-8X22B-INSTRUCT and some Meta-Llama3 models compared to those that omit it (Prompts 4–6).

6.2 Error Analysis

We examine the error rate of each model for different hallucination categories as well as highlight the proportion of errors caused by the models producing incorrect labels. The results are averaged across all prompts, as detailed in Figures 1 and 2.

The error rate seems to fluctuate across different hallucination categories, but without any strong or discernible patterns. We also find that a high error rate may be a result of the the number of invalid outputs (i.e., not $hyp1$ nor $hyp2$, nor any synonyms that correspond to either label) produced by some model. We notice this largely in MIXTRAL-8X22B-INSTRUCT, but to a lesser degree in GPT-SW3-20B-INSTRUCT, MIXTRAL-8X7B, and the two fairly small EuroLLM variants (respectively).

Notably, the Mixtral family tends to generate output claiming that both or neither hypotheses are hallucinations. Similarly, GPT-SW3 models dis-

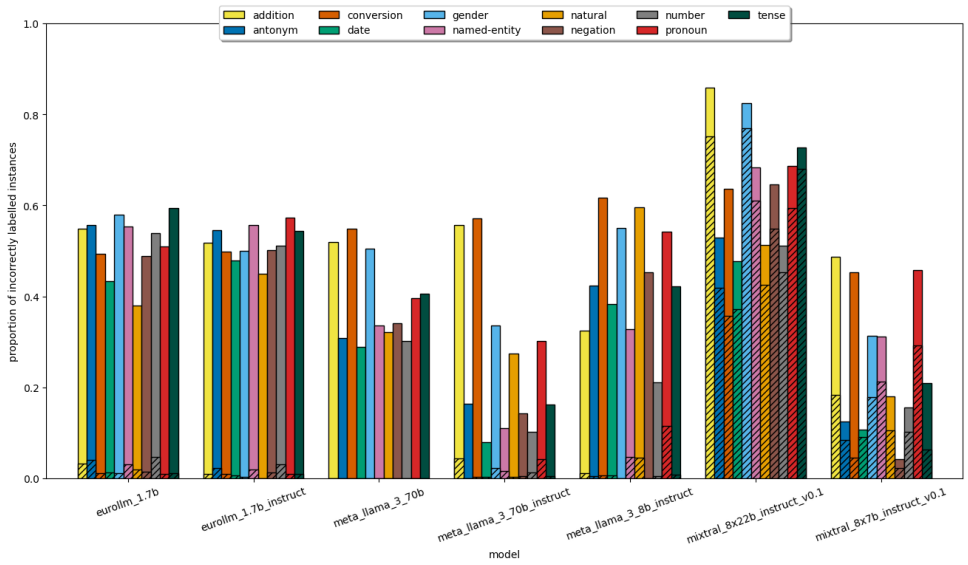

Figure 2: The average proportion of incorrectly labeled source-hyp **translation pairs** (averaged over all prompts and prompt and data language combinations) filtered by hallucination category. Here, the hatch represents the proportion of outputs that were invalid (i.e. falling outside $\{hyp1, hyp2\}$).

play a habit of returning a near-identical phrase or label for every instance. For example, GPT-SW3-20B tends to detect the $hyp1$ label for nearly every sentence pair, whereas the instruction-tuned variant has a higher error rate caused by invalid outputs, as it tends to, for some prompts, almost always output a phrase indicating its inability to perform the task (e.g.,*"It is hard to say without more context."*[3]). It is unclear why this model tends to converge on near-identical outputs, though it could relate to the type of data used during instruction tuning. Invalid outputs from the EuroLLM models, on the other hand, occur when the models start to translate or paraphrase the source sentence instead of performing the detection task at hand, although that is not surprising given their small size. It is worth noting that the NLI models' labels are determined by the entailment probabilities, which makes them immune to producing invalid labels, unlike the LLMs.

7 Conclusion

We have presented a suite of experiments to investigate the capabilities of open-access LLMs for detecting hallucinations, as defined in the HalluciGen task (Karlgren et al., 2024; Dürlich et al., 2024). The strongest models, MIXTRAL-8X7B-INSTRUCT and META-LLAMA-3-70B-INSTRUCT, perform consistently well across all languages and scenarios, suggesting that LLMs are appropriate for this task. The strong performance of the considerably smaller NLI models suggests that LLM-based detectors are not the only viable option.

We analyse the effect of four different factors: target language, model size, instruction-tuning and prompt – and find that none of them can be used as a straightforward predictor of model performance on this task. Our controlled experiments indicate that: **(i)** models perform consistently across languages, with a slight advantage for English; **(ii)** the impact of model size differs between model families; **(iii)** instruction-tuning has a clear positive effect only for the largest model; **(iv)** English prompts generally yield the best overall performance, while including the term "hallucination" in the prompt has a partially negative impact; and **(v)** for some models, a high error rate can be traced to the proportion of invalid outputs. We acknowledge the need for further investigation of these effects by systematically varying one factor at a time across different models.

In future work, we aim to explore whether LLMs may be used to *generate* datasets for training and evaluating hallucination detectors and apply these in a cross-model evaluation setting. In addition, given the relatively strong performance of NLI models in our experiments, it may be worth investigating whether other pre-existing techniques and metrics can be useful for detecting intrinsic hallucinations, including standard evaluation metrics for translation, paraphrasing and summarisation.

[3]In Swedish: "Det är svårt att säga utan mer sammanhang."

Acknowledgments

This work has been partially supported by the Swedish Research Council (grant number 2022-02909) and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10039436 [Utter]). We gratefully acknowledge EuroHPC JU (eurohpc-ju.europa.eu) for providing computing resources of the HPC system Leonardo Booster, hosted by the Interuniversity Consortium for Automatic Computing in North Eastern Italy. We thank the anonymous reviewers for their helpful suggestions.

Limitations

Owing to the very large and constantly expanding set of available LLMs and the numerous ways in which to prompt them, it is infeasible to conduct exhaustive prompt exploration experiments. In a similar vein, it is infeasible to explore all possible values for the generation parameters described in Section 4.2; though we selected values that should be broadly suitable, we did not optimise these for individual models. Nevertheless, we hope that our work provides insights into the suitability of LLMs as hallucination detectors, as indicated by their performance on the hallucination detection task.

When commenting on the presence of target languages in model pre-training data or the tasks included in instruction-tuning, we are reliant on information provided by the model developers in the form of academic papers, reports, and blog posts. Whilst these aspects are well documented for the EuroLLM and GPT-SW3 models, in the case of other models (e.g. Llama3 and Mixtral) this information may be incomplete or missing. Where such information is not provided, it is difficult to draw conclusions about the effects of different factors on model performance for any downstream task.

Additionally, two main limitations exist for the hallucination categories labels: (a) they suffer from class imbalance; and (b) they do not take into account that some samples could fall into multiple categories.

Our datasets focus only on a small set of high-resourced languages: English and Swedish for paraphrase and the English-French and English-German pairs for translation. Furthermore, a number of hallucination examples were constructed manually and may not accurately reflect real-world intrinsic hallucinations. Future work should look to reduce the English-centric nature of the datasets and expand the task to include a range of high, medium, and low-resource languages with exclusive focus on naturally occurring intrinsic hallucinations.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. pages 8137–8153, Singapore. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

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 Comput. Surv.*, 55(12).

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,

Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 288–298, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Haoqiang Kang, Terra Blevins, and Luke S. Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *ArXiv*, abs/2402.10496.

Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Liane Guillou, Joakim Nivre, Magnus Sahlgren, Aarne Talman, and Shorouq Zahra. 2024. Overview of eloquent 2024—shared tasks for evaluating generative language model quality. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 53–72, Cham. Springer Nature Switzerland.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint*, (2303.08896).

Mohamed El Marouani, Tarik Boudaa, and Nourddine Enneya. 2020. Machine translation evaluation using textual entailment for arabic. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–5.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Dan Saattrup Nielsen. 2022. Scandinli: Natural language inference for the scandinavian languages. `https://github.com/alexandrainst/ScandiNLI`.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305, Suntec, Singapore. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Hallucination Examples

Table 4 presents examples of hallucinated hypotheses for the paraphrase scenario for each hallucination category.

Type	Source	Hallucination
Addition	We struggle with water on a daily basis in the Netherlands - in the polders, the delta where the Meuse, the Rhine and the Scheldt flow into the sea.	In the Netherlands, we struggle with water on a daily basis because of the Meuse, Rhine, Scheldt, Noord, Voer and Dieze
Named-Entity	The fact is that a key omission from the proposals on agricultural policy in Agenda 2000 is a chapter on renewable energy.	Agenda 2030 does not include a chapter on renewable energy.
Number	The European Commission proposes that this information should enter into force within a period of three years from 1 July 1998.	The EU wants this information to enter into force in thirty years.
Conversion	In addition to these losses, there were also significant losses in terms of infrastructures, totalling approximately EUR 15 million.	There were losses in the amount of approximately 15 million dollars.
Date	In 1998, 1 700 000 net jobs were created in Europe, and although I admit that the employment situation is far from ideal, it has improved.	In 1700 there were 1 998 000 net jobs created in Europe.
Gender	Madam President, I am speaking on behalf of our colleague, Mr Francis Decourrière, who drafted one of the motions for a resolution.	One of the motions for a resolution was drafted by Mrs Francis Decourrière.
Pronoun	We have done so: on 5 February we published an extremely detailed press release dealing with the questions you have raised.	We published a press release that dealt with the questions we raised.
Antonym	The population has declined in some 210 of the 280 municipalities in Sweden, mainly in inland central and northern Sweden.	In the majority of Sweden's 280 municipalities, the population has gone up.
Tense	For the latter, the initial birth of several operators is now giving way to the reconcentration of the sector in the hands of a single company.	Several operators have given way to the reconcentration of the sector in the hands of one company.
Negation	The draft agenda as drawn up by the Conference of Presidents pursuant to Rule 95 of the Rules of Procedure has been distributed.	The Conference of Presidents hasn't distributed the draft agenda.
Natural	Amendment No 1 in the French version deletes illegal immigration and Amendment No 4 omits the expression 'police authorities'.	The French version excludes the expression'police authorities'.

Table 4: Examples of hallucination categories for the paraphrase task.

B Hallucination Statistics

Table 5 presents the frequency of each hallucination category for each language or language pair in the paraphrasing and machine translation hallucination detection scenarios, respectively. The data is first reported by (Dürlich et al., 2024).

Language	Scenario	Addition	Antonym	Date	Gender	Named Entity	Negation	Number	Pronoun	Tense	Conversion	Natural
en	PG	11	16	5	3	9	14	9	11	4	3	33
sv		42	11	–	3	15	12	9	1	5	1	20
en-fr	MT	10	–	24	–	33	–	33	–	–	–	–
fr-en		9	13	4	12	12	12	13	–	12	13	–
en-de		10	16	14	–	15	–	13	16	–	–	16
de-en		10	10	7	11	10	10	10	–	10	11	11

Table 5: Frequency statistics of each hallucination category across the different scenarios and languages.

C Non-English Prompts

Table 6 presents all non-English prompts used.

Prompt Name	Prompt
	Swedish paraphrase - Swedish prompt
Prompt 1	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en hallucination av den ursprungliga meningen. En hallucination innebär att hypotesen inte logiskt stöds av källan.
Prompt 2	Du är en AI-granskare specialiserad på parafrasdetektion. Din uppgift är följande: Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en hallucination av den ursprungliga meningen. En hallucination innebär att hypotesen inte logiskt stöds av källan.
Prompt 3	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en hallucination av den ursprungliga meningen.
Prompt 4	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som motsäger den ursprungliga meningen.
Prompt 5	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som stöder den ursprungliga meningen.
Prompt 6	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en dålig parafras av den ursprungliga meningen.
	Translation fr-en - French prompt
Prompt 1	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses est une hallucination de src. Une hallucination se manifeste quand l'original ne confirme pas logiquement l'hypothèse.
Prompt 2	Tu es un réviseur de traduction automatique IA. Ta tâche est la suivante: Tu reçois une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2). Décide laquelle des hypothèses est une hallucination de src. Une hallucination se manifeste quand l'original ne confirme pas logiquement l'hypothèse.
Prompt 3	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses est une hallucination de src.
Prompt 4	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses contredit src.
Prompt 5	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses confirme src.
Prompt 6	Tu es un réviseur IA avec une spécialisation en traduction automatique. Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses est une mauvaise traduction de src.
	Translation de-en - German prompt
Prompt 1	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen halluziniert ist. Eine Halluzination tritt auf, wenn die Hypothese das Original (src) nicht logisch unterstützt.
Prompt 2	Du bist ein KI-Prüfer für maschinelle Übersetzung. Deine Aufgabe ist die folgende: Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen halluziniert ist. Eine Halluzination tritt auf, wenn die Hypothese das Original (src) nicht logisch unterstützt.
Prompt 3	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen halluziniert ist.
Prompt 4	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen src widerspricht.
Prompt 5	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen src unterstützt.
Prompt 6	Du bist ein KI-Prüfer mit Fachkenntnissen in maschineller Übersetzung. Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen eine schlechte Übersetzung von src ist.

Table 6: Prompt formulations tested in Swedish, French and German.

D Label Post-Processing

The tested models usually return one of the two expected labels verbatim ($hyp1$ or $hyp2$), but some models tend to return the label in a different phrasing. For this reason, we first check if the generated model output contains any of these variations:

- "1" or "2"
- "hyp 1" or "hyp 2" (including whitespace)
- "hypotes 1" or "hypotes 2"
- "hypothèse 1" or "hypothèse 2"

- "hypothese 1" or "hypothese 2"

If the model output contains only one label (in whatever variation), we extract that as the label. If the generated output contains both labels, we consider the output invalid and return an empty label. If none of the variations above are present, we expand the list of variations to cover the different languages in which the models are prompted:

- "hyp1" or "hyp2" (no whitespace)
- "hypothesis1" or "hypothesis12"
- "first or "second"
- "första or "andra"
- "erste" or "zweite"
- "première/premier" or "deuxième"
- "hypotes1" or "hypotes2"
- "hypothèse1" or "hypothèse2"
- "hypothese1" or "hypothese2"

As explained in Section 4.2, Prompt 5 is formulated in such a way that the task is reversed; we prompt the model to output a label for the hypothesis that *supports* the source. For this reason, and for this particular prompt only, the label is flipped from $hyp1$ to $hyp2$ and vice versa unless the model produces an empty label (in which case the label is kept as is).

E Model repositories

Family	Variant	Repository	Version
Llama-3	META-LLAMA-3-8B-INSTRUCT	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct	3.0
	META-LLAMA-3-70B-INSTRUCT	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct	3.0
	META-LLAMA-3-70B	https://huggingface.co/meta-llama/Meta-Llama-3-70B	3.0
Mixtral	MIXTRAL-8X7B-INSTRUCT	mistralai/Mixtral-8x7B-Instruct-v0.1	v0.1
	MIXTRAL-8X22B-INSTRUCT	https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1	v0.1
EuroLLM	EUROLLM-1.7B	https://huggingface.co/utter-project/EuroLLM-1.7B	-
	EUROLLM-1.7B-INSTRUCT	https://huggingface.co/utter-project/EuroLLM-1.7B-Instruct	-
GPT-SW3	GPT-SW3-20B-INSTRUCT	https://huggingface.co/AI-Sweden-Models/gpt-sw3-20b-instruct	-
	GPT-SW3-20B	https://huggingface.co/AI-Sweden-Models/gpt-sw3-20b	-
	GPT-SW3-40B	https://huggingface.co/AI-Sweden-Models/gpt-sw3-40b	-

F NLI Baselines Details

To determine which of the two hypotheses ($hyp1$, $hyp2$) contains a hallucination, we predict "entailment" (E) and "not_entailment" (NE) class scores between the source sentence and each one of the hypotheses. We then choose the hallucination based on which one or more hypotheses

- If **E** $>$ **NE** for one hypothesis and **E** $<$ **NE** for the other, we choose the one with **E** $<$ **NE**.
- If **E** $>$ **NE** for both hypotheses, we choose the one with the lowest **E** score.
- If **E** $<$ **NE** for both hypotheses, we choose the one with the highest **NE** score.

G Compute Environment and Efficiency

The experiments were performed on Leonardo Booster[4], equipped with NVidia A100 SXM6 64GB GPUs with a single 32-core Intel Ice Lake CPU. Model inference is performed sequentially (in other words, without batching) for each sample, using the Accelerate library from Huggingface.[5] Table 7 presents the number of GPUs used for loading each model, as well as execution time for performing inference on a single model input.

[4] https://leonardo-supercomputer.cineca.eu/hpc-system/
[5] https://pypi.org/project/accelerate

Model name	Number of GPUs	Inference time per sample (sec)
META-LLAMA-3-8B-INSTRUCT	2	7.01
META-LLAMA-3-70B	4	11.44
META-LLAMA-3-70B-INSTRUCT	4	14.77
MIXTRAL-8X7B-INSTRUCT	2	15.13
MIXTRAL-8X22B-INSTRUCT	4	23.10
EUROLLM-1.7B	1	18.34
EUROLLM-1.7B-INSTRUCT	1	19.94
GPT-SW3-20B	1	14.45
GPT-SW3-20B-INSTRUCT	1	12.46
GPT-SW3-40B	3	13.02

Table 7: Number of GPUs used for loading each model, as well as execution time for performing inference on one input.

H Annotation Guidelines: Paraphrase Hallucinations

Task: Your task is to mark each sentence as hallucination (H) or not hallucination (NH).

Definition of hallucination for this task: Given a src and a generated hypothesis hyp in the context of paraphrasing, we ask the question: is hyp supported by the src? If yes, then hyp is marked as not hallucination (NH). If no, then hyp is marked as hallucination (H).

A hypothesis *supports* the source when:

- The overall semantics of the source are preserved, but some minor details are missing

A hypothesis *does not support* the source when:

- New information, i.e. information that was not present in the source and could not be deduced from the source, is added
- It contains nonsensical information (when the source does not)
- It misrepresents the semantic relationships in the source (i.e. a bad paraphrase)

Example:

Src	Stockholm is the capital of Sweden and is located on the East coast
Hyp (NH)	1) Stockholm, situated on the East coast, serves as the capital of Sweden 2) Stockholm is situated on the East coast
Hyp (H)	Stockholm is the capital of Denmark

The annotators for the paraphrase data are the authors of this paper, and all are fluent speakers of English and/or Swedish.