# Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models

**Anonymous ACL submission**

## Abstract

Recent advancements in massively multilingual machine translation systems have significantly enhanced translation accuracy; however, even the best performing systems still generate hallucinations, severely impacting user trust. Detecting hallucinations in Machine Translation (MT) remains a critical challenge, particularly since existing methods excel with High-Resource Languages (HRLs) but exhibit substantial limitations when applied to Low-Resource Languages (LRLs). This paper evaluates hallucination detection approaches using Large Language Models (LLMs) and semantic similarity within massively multilingual embeddings. Our study spans 16 language directions, covering HRLs, LRLs, with diverse scripts. We find that the choice of model is essential for performance. On average, for HRLs, `Llama3-70B` outperforms the previous state of the art by as much as 0.16 MCC (Matthews Correlation Coefficient). However, for LRLs we observe that `Claude Sonnet` outperforms other LLMs on average by 0.03 MCC. The key takeaway from our study is that LLMs can achieve performance comparable or even better than previously proposed models, despite not being explicitly trained for any machine translation task. However, their advantage is less significant for LRLs. [1]

## 1 Introduction

Text generation models have drastically improved in recent years especially with the capabilities of LLMs in producing realistic and fluent output. However, hallucination continues to undermine user trust, as it generates and propagates misinformation and sometimes nonsensical outputs (Agarwal et al., 2018; Xu et al., 2023a; Guerreiro et al., 2023b).

One practical way of reducing hallucination in MT is by building more robust models, especially for LRL which tend to exhibit significantly higher hallucination rates. There are several efforts on scaling MT models to LRLs, such as M2M-100 (Fan et al., 2020), NLLB-200 (Team et al., 2022), MADLAD-400 (Kudugunta et al., 2023) etc. Despite initiatives to minimize hallucinations during the MT process, issues still persists. Therefore, detecting hallucinations post-translation remains a critical alternative approach to ensure the reliability and trustworthiness of the translated content.

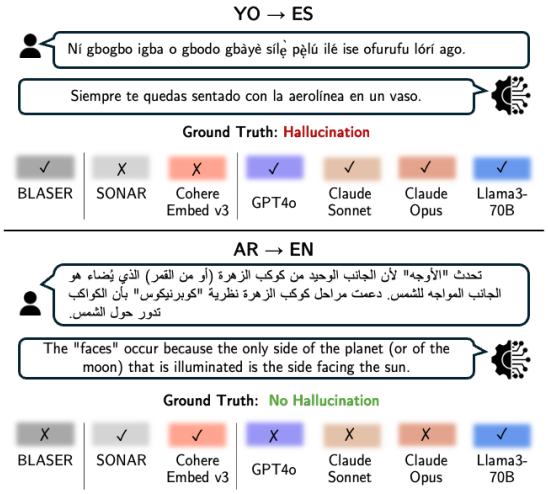


Figure 1: Illustration of how a selection of the evaluated methods perform from Yoruba to Spanish and from Arabic to English.

Previous work on post-translation evaluation has mainly focused on a single English-centric (`EN`) to a HRL direction, while studies including LRL remain limited (Raunak et al., 2022; Xu et al., 2023b). Recently, Dale et al. (2023) introduced *HalOmi*— a benchmark dataset for detecting hallucination in MT that includes `EN↔HRLs` (ten directions) and `EN↔LRLs` (six directions), as well as two non-English directions `HRL↔LRL`, including different scripts. `BLASER-QE` (Communication et al., 2023), the state-of-the-art (SOTA) hallucination detector, is reported as the top performer on the *HalOmi* benchmark. It calculates a translation

[1] Code will be released upon acceptance

quality score by evaluating the similarity between encoded source texts and machine-translated texts within the SONAR embedding space (Duquenne et al., 2023).

In this paper, we evaluate the performance of LLMs and embedding based methods as hallucination detectors, aiming to enhance performance in both HRLs and LRLs. To this end, we use the *HalOmi* benchmark dataset with a binary hallucination detection approach. For our evaluation, we include 14 methods: eight LLMs with different prompt variations, and four embedding spaces by computing the cosine similarity between source and translated texts.

We find that LLMs are highly effective for hallucination detection across both high and low resource languages, although the optimal model selection depends on specific contexts. For HRLs, on average across directions, the `Llama3-70B` model significantly surpasses the previous SOTA method, `BLASER-QE`, by 16 points. Moreover, embedding-based methods have also demonstrated superior performance over the current SOTA in high resource contexts. However, for LRLs, `Claude Sonnet` is the best performing model, improving previous methods by a smaller difference. More precisely, LLMs outperformed `BLASER-QE` in five out of eight LRL translation directions, including the non-English-centric ones.

Finally, our research makes the following primary contributions: First, we evaluate a wide range of LLMs for MT hallucination detection and establish that LLMs, despite not being explicitly trained for the task, are competitive and *greatly* outperform even the previous SOTA for HRLs. Second, large multilingual embedding spaces improve upon previously proposed methods and show that they remain competitive for HRLs, but struggle for LRLs. Third, we establish a new SOTA for 13 of the 16 languages that we evaluate on, including high and low resource languages. Surpassing the previous SOTA, which was explicitly trained for the task, on average by 2 MCC points.

## 2 Experimental setup

### 2.1 Quality assessment of the dataset

We evaluated our methods on the *HalOmi* dataset. A first dataset filtration involved selecting only natural translations, without perturbations, as findings from perturbed data may not be applicable to the detection of natural hallucinations (Dale et al., 2023).

The validation and test split was decided based on the translation direction. For the validation set, we selected the two translation directions `DE`↔`EN`, which encompasses `301` sentences. This choice was made as extensive resources and established benchmarks are available for this language pair (Guerreiro et al., 2023a), with the expectation that the models would exhibit generalizability to less frequently used language pairs. For the test set, the other 16 pairs were used: more precisely, it includes four pairs with English and a HRL (`EN`↔`AR` , `EN`↔`ZH`, `EN`↔`RU`, `EN`↔`ES`), three pairs with English and a LRL (`EN`↔`KS`, `EN`↔`MN`, and `EN`↔`YO`), and one non-English HRL-LRL pair (`ES`↔`YO`). The test set includes `2,558` sentence pairs. This test set excludes six sentence pairs that were removed due to sensitive content flagged and filtered out by LLMs. A more detailed description of the dataset is available in Appendix B.

### 2.2 Hallucination detection setting

We consider two settings: (1) **Severity ranking** introduced by the authors of *HalOmi*. (2) **Binary detection**—a new setting we added due to data imbalance and ease of evaluation.

**Severity ranking** the classification of hallucinations was based on four severity levels: *No Hallucination*, *Small Hallucination*, *Partial Hallucination*, and *Full Hallucination*. This fine-grained categorization aimed to capture the nuances in the extent and impact of hallucinations on the translated output. We use this setting only as **ablation study** in Appendix C., both for consistency with the *HalOmi* benchmark, but also to assess the relevance of our binary detection approach.

**Binary detection** In this setting, all three instances of hallucinations were labelled as *Hallucination*, regardless of their severity. We also change the way the evaluation was done in *HalOmi*, with an appropriate prompt (Appendix D), and threshold calculation for binary classification for embeddings cosine similarity, see subsection 2.4. The primary reason for choosing this setting is the significant class imbalance in *HalOmi*, largely due to the scarcity of hallucinations across different severity levels. Some translation directions have particularly imbalanced data, for example `EN`→`RU`, with the following distribution: out of 148 sentence pairs, we have 141 *No Hallucination* (96.6%), 1 *Small* (0.68%), 2 *Partial* (1.4%), and 4 *Full* (2.8%). High class imbalance can affect the ability of model

to perform well (Prusa et al., 2016; Sordo and Zeng, 2005; Fernández et al., 2013).

### 2.3 LLMs for hallucination detection

We assessed the performances of eight LLMs, mixing capabilities models across LLMs families. We evaluate OpenAI's `GPT4-turbo` and `GPT4o`; Cohere's `Command R` and `Command R+`; Mistral's `Mistral-8x22b`; Anthropic's `Claude Sonnet` and `Claude Opus` and Meta's `Llama3-70B`.[2] More details about the selection are in subsection E.2.

First, we built our prompt design by differentiated system and user prompts for better results (Kong et al., 2024). The system prompt contained the task description, and optionally, the inclusion of Chain-of-Thought (CoT), while the user prompt contained, for each sentence pair, the source text and MT text, as well as a direct hallucination classification question.

We derived the task description prompts from the *Evaluate Hallucination* and *Evaluate Coherence in the Summarization Task* prompts in *G-Eval* (Liu et al., 2023). The CoT prompts were inspired by *Evaluation Steps* from *G-Eval*, and by the human annotation guidelines and severity level definitions from *HalOmi*. All prompts are available Appendix D. More details about the chosen hyperparameters with LLMs can be found in Appendix E.

We determined the optimal prompts for each model using the `DE↔EN` validation set, evaluating three prompts and two CoT proposals for binary detection. The best prompt for each model was selected based on the average MCC across both translation directions. The MCC was chosen as the primary metric for binary detection due to its superiority in providing a single, easily interpretable value between -1 and +1. This value encapsulates the model's performance for the confusion matrix scores, making it more robust to class imbalance.

### 2.4 Embeddings

We assessed the performance of three LLM-related embedding spaces: OpenAI's `text-embedding-3-large`, Cohere's `Embed v3`, and Mistral's `mistral-embed`. Additionally, we included `SONAR`, the multilingual embedding space used as the base for `BLASER-QE`. Specifically, we calculated the cosine distance between embeddings of the source text and the machine-translated text. This approach draws on previous studies showing that hallucinated translations tend to have embeddings that are significantly distanced from those of the source text (Dale et al., 2022).

We binarised the cosine similarity scores of embeddings using an optimal threshold value determined from the validation set. This threshold, established by maximizing the F1-score from the precision-recall curve, was then applied to the test set for binary hallucination detection across all language pairs. Each embedding space was independently processed to maintain the integrity of the evaluation.

[2] `GPT3.5`, `Mistral Large` and `Llama3-8B` were initially taken into account, but were excluded due to poor task understanding.

| | | | EMBEDDINGS | | | | LLMs | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BLASER-QE | SONAR | GPT | Cohere | Mistral | GPT4-turbo | GPT4o | Command-R | Command-R Plus | Mistral 8x22b | Claude Sonnet | Claude Opus | Llama3-70B |
| HRLs | EN → AR | 0.41 | 0.40 | 0.35 | 0.40 | 0.29 | 0.27 | 0.38 | 0.43 | 0.14 | 0.25 | 0.39 | 0.39 | 0.39 |
| | AR → EN | 0.69 | 0.76 | 0.77 | 0.69 | 0.70 | 0.47 | 0.52 | 0.56 | 0.57 | 0.62 | 0.58 | 0.59 | 0.77 |
| | EN → RU | 0.13 | 0.17 | 0.20 | 0.20 | 0.24 | 0.27 | 0.34 | 0.34 | -0.03 | 0.59 | 0.52 | 0.26 | 0.53 |
| | RU → EN | 0.64 | 0.75 | 0.61 | 0.67 | 0.67 | 0.34 | 0.43 | 0.39 | 0.50 | 0.53 | 0.43 | 0.48 | 0.72 |
| | EN → ES | 0.61 | 0.41 | 0.66 | 0.66 | 0.70 | 0.61 | 0.53 | 0.52 | 0.55 | 0.55 | 0.59 | 0.59 | 0.66 |
| | ES → EN | 0.51 | 0.50 | 0.58 | 0.51 | 0.47 | 0.55 | 0.50 | 0.59 | 0.63 | 0.55 | 0.64 | 0.58 | 0.68 |
| | EN → ZH | 0.31 | 0.51 | 0.78 | 0.71 | 0.63 | 0.53 | 0.35 | 0.57 | 0.36 | 0.72 | 0.56 | 0.63 | 0.78 |
| | ZH → EN | 0.45 | 0.65 | 0.60 | 0.55 | 0.60 | 0.39 | 0.41 | 0.43 | 0.41 | 0.45 | 0.46 | 0.39 | 0.51 |
| | **HRLs AVG** ±STD | **0.47** ±0.19 | **0.52** ±0.20 | **0.57** ±0.20 | **0.55** ±0.18 | **0.54** ±0.18 | **0.43** ±0.13 | **0.43** ±0.08 | **0.48** ±0.09 | **0.39** ±0.23 | **0.53** ±0.14 | **0.52** ±0.09 | **0.49** ±0.13 | **0.63** ±0.14 |
| LRLs | EN → KA | 0.39 | 0.44 | 0.19 | 0.42 | 0.39 | 0.53 | 0.47 | 0.30 | 0.34 | 0.49 | 0.49 | 0.52 | 0.39 |
| | KA → EN | 0.37 | 0.29 | 0.14 | 0.02 | 0.06 | 0.06 | 0.12 | 0.14 | 0.03 | 0.15 | 0.36 | 0.21 | 0.21 |
| | EN → YO | 0.43 | 0.31 | 0.21 | 0.25 | 0.26 | 0.37 | 0.38 | 0.25 | 0.35 | 0.26 | 0.47 | 0.36 | 0.33 |
| | YO → EN | 0.40 | 0.11 | 0.17 | 0.20 | 0.17 | 0.16 | 0.22 | 0.06 | 0.20 | 0.15 | 0.25 | 0.20 | 0.26 |
| | EN → MN | 0.37 | 0.21 | -0.07 | 0.00 | 0.03 | 0.20 | 0.14 | -0.01 | 0.00 | 0.06 | 0.07 | 0.02 | 0.11 |
| | MN → EN | 0.07 | 0.15 | 0.02 | 0.19 | 0.12 | 0.12 | 0.30 | 0.19 | 0.18 | 0.09 | 0.21 | 0.16 | 0.16 |
| HRL-LRL | ES → YO | 0.17 | 0.16 | 0.10 | 0.15 | 0.22 | 0.22 | 0.22 | 0.20 | 0.19 | 0.19 | 0.35 | 0.32 | 0.22 |
| | YO → ES | 0.06 | -0.07 | 0.04 | 0.09 | -0.10 | 0.12 | 0.25 | 0.09 | 0.12 | 0.21 | 0.15 | 0.24 | 0.12 |
| | **LRLs AVG** ±STD | **0.28** ±0.16 | **0.20** ±0.15 | **0.10** ±0.10 | **0.16** ±0.13 | **0.14** ±0.15 | **0.22** ±0.16 | **0.26** ±0.12 | **0.15** ±0.11 | **0.18** ±0.13 | **0.20** ±0.13 | **0.29** ±0.15 | **0.25** ±0.15 | **0.22** ±0.10 |
| | **OVERALL AVG** ±STD | **0.38** ±0.19 | **0.36** ±0.24 | **0.33** ±0.29 | **0.36** ±0.25 | **0.34** ±0.26 | **0.33** ±0.18 | **0.35** ±0.13 | **0.32** ±0.19 | **0.28** ±0.21 | **0.37** ±0.22 | **0.41** ±0.17 | **0.37** ±0.18 | **0.43** ±0.24 |

Figure 2: MCC scores for hallucination binary detection across 16 translation directions per method.

## 3 Results

**LLMs are the new SOTA for hallucination detection** The results in Figure 2 demonstrate that LLMs have the best overall performance across languages for binary hallucination detection. Specifically, `Llama3-70B` surpasses the previous best performing model, `BLASER-QE`, by +5 points, with an MCC of 0.43. For HRLs, 10 out of 12 evaluated methods outperform `BLASER-QE` (0.46), with `Llama3-70B` greatly improving over the baseline by 16 points (0.63). Notably, the results show that the choice of LLM should rely on the resource

level; as for LRLs, Claude Sonnet achieves the highest average MCC. However, GPT4o was the more robust LLM across all languages, with the lowest standard deviation. Finally, for 13 out of the 16 evaluated translation directions, the evaluated methods outperform BLASER-QE, with the exception of KS→EN, YO→EN and EN→MNI. Our findings on LLMs' superior hallucination detection capabilities align with prior research on their effectiveness in MT quality assessment (Kocmi and Federmann, 2023).

**Embedding-based hallucination detectors remain competitive for HRLs** For HRLs, simple embedding-based methods display competitive capabilities, outperforming more sophisticated models in five out of eight translation directions. For instance, although BLASER-QE is a more advanced model based on SONAR, SONAR exhibits comparable or superior performances in most HRLs directions. This suggests that the effectiveness of these methods may be highly sensitive on their training data, and hence to the resource level, as we observe SOTA performances for HRLs and suboptimal results for LRLs. Additionally, the embeddings' performance may be highly dependent on the threshold chosen using the EN↔DE validation set, generalizing well for HRLs but not for LRLs.

**LLMs' contrastive performances across LRLs** First, while Llama3-70B obtains the best performance overall, it was outperformed in most translation directions, especially in LRL. This result reveals a HRLs-centric approach of the model but also concludes that there is not one-LLM fits all resource levels. Secondly, for LRLs, models such as Sonnet, Opus, GPT4o, and Mistral —in order of decreasing performances, achieve higher scores, supporting the feasibility of employing LLMs in settings encompassing a wide range of languages. These results should be contrasted with a wide difference of hallucination distribution across resource levels, for example with the MN→EN direction which only has 28% *No hallucination* sentence pairs. More precisely, Sonnet and BLASER-QE perform on par for LRL, with the particularity that BLASER-QE has a significantly higher rate of false negatives, while Sonnet maintains a more balanced ratio of false positives to negatives. Moreover, BLASER-QE performs well in translations from English and comparably to Sonnet in translations to English, but falls short in non-English-centric translations, which follows the same trends as previously reported models in (Dale et al., 2023). Figure 14 provides a more detailed view of these performance metrics.

**Embeddings are high performers for non-Latin scripts, while LLMs can generalise to non-English centric translations** For HRLs→EN directions with source scripts different than Latin (AR,RU,ZH), embeddings are the best performers, suggesting high capabilities with cross-script transfer learning. These observations align with the findings of Hada et al. (2023), who report decreased performance for non-Latin scripts in LLM-based evaluators. In the two non-English centric translation directions (ES↔YO), Opus outperforms by far both BLASER-QE (0.11) and the best embedding Mistral (0.12), with a score of 0.28. Unlike the overall LRLs trends, Opus outperforms Sonnet for this direction pair: this can suggest that the advanced analytical capabilities of LLMs can generate improved results even in scenarios with limited relevant training data. Remarkably, in the YO→ES translation direction, six out of our fourteen methods and BLASER-QE exhibit scores close to random guessing (within the [–1, +1] range). This observation underscores the pressing need for enhanced capabilities in detecting hallucinations in non-English-centric translation settings. Figure 1 presents two examples that highlight the challenges faced by LLMs when dealing with non-Latin scripts, with the exception of Llama3-70B. Additionally, it illustrates how embeddings may struggle with reasoning capabilities in non-English centric contexts.

## 4 Conclusion

In this work, we demonstrates that LLMs and embedding semantic similarity are highly effective for hallucination detection in machine translation, with LLMs establishing a new state-of-the-art performance across both high and low-resource languages. Our findings suggest that the optimal model selection depends on specific contexts, such as resource level, script, and translation direction. Our study highlight the need for further research to enhance hallucination detection capabilities, particularly in low-resource and non-English-centric translation settings.

## Limitations

Despite the promising results obtained by LLMs and embedding-based methods in our evaluation, there are certain limitations that should be noted.

First, the dataset shows distribution imbalance across translation directions, with different trends for high and low resource languages, even after binarisation (see Appendix B): The HRLs show a pronounced data imbalance towards *No Hallucination* labels, with distribution between $79\%$ and $94\%$. Moreover, for LRLs, there's a broader interval, from $28\%$ to $85\%$. This imbalance often results in models that classify translations as *No hallucination* being more frequently correct for HRLs than for LRLs, thereby introducing a bias into the binary evaluation. Moreover, the translation direction display a qualitative bias, as shown subsection B.3: HRLs and LRLs don't have the same selection distribution which display a potential bias towards hallucination. Future dataset improvements should prioritize larger, more diverse samples, non-Latin scripts, and non-English centric translations. Using consistent source text across languages and balancing hallucination severity levels would enable more sophisticated methods, improve generalizability, and allow for a fair evaluation of models' hallucination detection capabilities.

The validation set used to identify the optimal threshold for non-LLM methods and the best prompt for LLMs only included EN↔DE translations. To improve parameter optimization and generalization across various translation directions, especially for low-resource languages (LRLs), cross-validation is recommended for future research, as suggested by Dale et al. (2023) and initially planned for our study. However, financial constraints associated with benchmarking non-open source models prevented the implementation of this approach. Future work should focus on developing novel approaches that perform well on well-studied high-resource languages (HRLs) while generalizing effectively to LRLs, assessing robustness, or exploring alternative methods to address this challenge within the limitations of dataset size.

Finally, for benchmarking purposes, only the previous state-of-the-art (SOTA) was included for comparison against the newly evaluated methods. Therefore, for a more comprehensive analysis, it is recommended to include additional methods previously evaluated by *HalOmi*.

## References


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

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-Jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:36–50.

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

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations.

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. *Preprint*, arXiv:2010.11125.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:878–891.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Alberto Fernández, Victoria López, Mikel Galar, María José Del Jesus, and Francisco Herrera. 2013. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, 42:97–110.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023a. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

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.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2024*, pages 1051–1070.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023*, pages 193–203.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. *Preprint*, arXiv:2308.07702.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy

Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Joseph Prusa, Taghi M. Khoshgoftaar, and Naeem Seliya. 2016. The effect of dataset size on training tweet sentiment classifiers. *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, pages 96–102.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179.

Margarita Sordo and Qing Zeng. 2005. On sample size and classification accuracy: A performance comparison. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3745 LNBI:193–201.

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. *Preprint*, arXiv:2207.04672.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023a. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023b. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

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

# A Related Work

Significant advancements have been made in automatic machine translation evaluation, but these have predominantly focused on general translation errors. As a result, hallucinations are often overlooked, and evaluation scores may not reflect their impact due to their relatively low frequency compared to less severe errors like omissions (Guerreiro et al., 2023a).

Previous studies have demonstrated that sentence similarity measures between source and translated texts, using cross-lingual embeddings such as LASER (Heffernan et al., 2022) and LaBSE (Feng et al., 2022), can effectively identify severe hallucinations (Dale et al., 2022). However, the recently introduced *Halomi* dataset, *A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation* (Dale et al., 2023), which expands to include LRLs and non-English-centric translation directions, reveals the limitations of embedding semantic similarity methods primarily with LRLs. Conversely, the BLASER model (Communication et al., 2023)—utilizing the SONAR embedding space (Duquenne et al., 2023)—demonstrates greater robustness across language resources, establishing it as the latest state-of-the-art. This model notably improves performance in LRLs compared to previous methods, yet it still shows deficiencies in some non-English-centric directions.

Recent works have underlined the capabilities of LLMs in multilingual MT evaluation, demonstrating strong performances across various languages, although discrepancies are noted in LRLs (Zhu et al., 2023; Xu et al., 2023b). *G-Eval* (Liu et al., 2023) introduces a robust prompting framework for hallucination detection and demonstrates that LLMs can be used as automatic metrics to generate a single quality score. Furthermore, Kocmi and Federmann (2023) showed that LLMs, when appropriately prompted, can assess the quality of machine-generated translations, achieving state-of-the-art performance in system-level quality evaluation. Moreover, Fernandes et al. (2023) pioneered the evaluation of LLMs for MT tasks in LRLs using a new prompting technique, although its focus is primarily on broader translation errors rather than specifically on hallucination detection.

# B Dataset description

## B.1 Language acronyms mapping

The languages acronyms follow this mapping throughout the paper: Arabic (AR), Chinese (ZH), English (EN), German (DE), Kashmiri (KA), Manipuri (MN), Russian (RU), Spanish (ES), and Yoruba (YO).

## B.2 Hallucination distribution

### B.2.1 Distribution of Hallucination in the severity ranking framework

| Direction | **Total** | *1 No* | *2 Small* | *3 Partial* | *4 Full* |
|---|---|---|---|---|---|
| DE→EN | **155** | 140 | 2 | 2 | 11 |
| | | *90.32%* | *1.29%* | *1.29%* | *7.10%* |
| EN→DE | **146** | 132 | 3 | 2 | 9 |
| | | *90.41%* | *2.05%* | *1.37%* | *6.16%* |
| **Total** | **301** | 272 | 5 | 4 | 20 |
| | | 68.25% | 1.25% | 1.00% | 5.01% |

| Direction | **Total** | *1 No* | *2 Small* | *3 Partial* | *4 Full* |
|---|---|---|---|---|---|
| EN→AR | **144** | 136 | 2 | 2 | 4 |
| | | *94.44%* | *1.39%* | *1.39%* | *2.78%* |
| AR→EN | **156** | 132 | 5 | 2 | 17 |
| | | *84.62%* | *3.21%* | *1.28%* | *10.90%* |
| EN→RU | **146** | 141 | 1 | 2 | 2 |
| | | *96.58%* | *0.68%* | *1.37%* | *1.37%* |
| RU→EN | **158** | 146 | 3 | 2 | 7 |
| | | *92.41%* | *1.90%* | *1.27%* | *4.43%* |
| EN→ES | **153** | 131 | 8 | 3 | 11 |
| | | *85.62%* | *5.23%* | *1.96%* | *7.19%* |
| ES→EN | **160** | 127 | 17 | 4 | 12 |
| | | *79.38%* | *10.63%* | *2.50%* | *7.50%* |
| EN→ZH | **160** | 131 | 5 | 4 | 20 |
| | | *81.88%* | *3.13%* | *2.50%* | *12.50%* |
| ZH→EN | **159** | 127 | 9 | 7 | 16 |
| | | *79.87%* | *5.66%* | *4.40%* | *10.06%* |
| EN→KA | **184** | 111 | 8 | 30 | 35 |
| | | *60.33%* | *4.35%* | *16.30%* | *19.02%* |
| KA→EN | **151** | 89 | 15 | 32 | 15 |
| | | *58.94%* | *9.93%* | *21.19%* | *9.93%* |
| EN→YO | **195** | 166 | 4 | 11 | 14 |
| | | *85.13%* | *2.05%* | *5.64%* | *7.18%* |
| YO→EN | **146** | 124 | 4 | 10 | 8 |
| | | *84.93%* | *2.74%* | *6.85%* | *5.48%* |
| EN→MN | **197** | 78 | 52 | 54 | 13 |
| | | *39.59%* | *26.40%* | *27.41%* | *6.60%* |
| MN→EN | **152** | 43 | 45 | 58 | 6 |
| | | *28.29%* | *29.61%* | *38.16%* | *3.95%* |
| ES→YO | **151** | 97 | 16 | 29 | 9 |
| | | *64.24%* | *10.60%* | *19.21%* | *5.96%* |
| YO→ES | **152** | 80 | 26 | 37 | 9 |
| | | *52.63%* | *17.11%* | *24.34%* | *5.92%* |
| **Total** | **2564** | 1859 | 220 | 287 | 198 |
| | | *72.47%* | *8.58%* | *11.19%* | *7.72%* |

Table 1: Although fine-grained severity ranking is advantageous for most applications, the rarity of occurrences within each hallucination category may lead to results that lack significance and generalizability due to constrained sample sizes. Notably, within the *HalOmi* dataset, 11 of the 18 language directions include fewer than five samples in at least one hallucination category. To address this limitation, we propose a shift toward binary hallucination detection, where all instances of hallucinations are classified as such, irrespective of their severity. This approach enhances the robustness of the analysis and the significance of results while still evaluating the model's ability to separate even *Small hallucination* (one word in a sentence) from *No hallucinations.*

### B.2.2 Distribution of Hallucination in the binary detection framework

| Direction | Total | *0 No Hallucination* | *1 Hallucination* |
|---|---|---|---|
| DEU→EN | **155** | 140 | 15 |
| | | *90.32%* | *9.68%* |
| EN→DE | **146** | 132 | 14 |
| | | *90.41%* | *10.00%* |
| **Total** | **301** | 272 | 29 |
| | | 68.25% | 31.75% |

Table 2: **Validation set distribution for binary detection, across translation directions, for HRLs and LRLS**

| Direction | Total | *0 No Hallucination* | *1 Hallucination* |
|---|---|---|---|
| EN→AR | **144** | 136 | 8 |
| | | *94.44%* | *5.56%* |
| AR→EN | **156** | 132 | 24 |
| | | *84.62%* | *15.38%* |
| EN→RU | **146** | 141 | 5 |
| | | *96.58%* | *3.42%* |
| RU→EN | **158** | 146 | 12 |
| | | *92.41%* | *7.59%* |
| EN→ES | **153** | 131 | 22 |
| | | *85.62%* | *14.38%* |
| ES→EN | **160** | 127 | 33 |
| | | *79.38%* | *20.63%* |
| EN→ZH | **160** | 131 | 29 |
| | | *81.88%* | *18.13%* |
| ZH→EN | **159** | 127 | 32 |
| | | *79.87%* | *20.13%* |
| EN→KA | **184** | 111 | 73 |
| | | *60.33%* | *39.67%* |
| KA→EN | **151** | 89 | 62 |
| | | *58.94%* | *41.06%* |
| EN→YO | **195** | 166 | 29 |
| | | *85.13%* | *14.87%* |
| YO→EN | **146** | 124 | 22 |
| | | *84.93%* | *15.07%* |
| EN→MN | **197** | 78 | 119 |
| | | *39.59%* | *60.41%* |
| MN→EN | **152** | 43 | 109 |
| | | *28.29%* | *71.71%* |
| ES→YO | **151** | 97 | 54 |
| | | *64.24%* | *35.76%* |
| YO→ES | **152** | 80 | 72 |
| | | *52.63%* | *47.37%* |
| **Total** | **2564** | 1859 | 705 |
| | | *72.47%* | *27.53%* |

Table 3: **Testing set distribution for binary detection, across translation directions, for HRLs and LRLS**

### B.3 Selection distribution

The selection information from the *HalOmi* dataset indicates the sampling strategy used to select sentence pairs for each translation direction and data source, which includes *uniform* sampling to maintain data diversity, *biased* sampling favoring potentially problematic translations based on detector quantiles, and *worst* sampling, according to the detectors to increase the likelihood of capturing hallucinations. A closer look at the selection distribution is available Figure 3

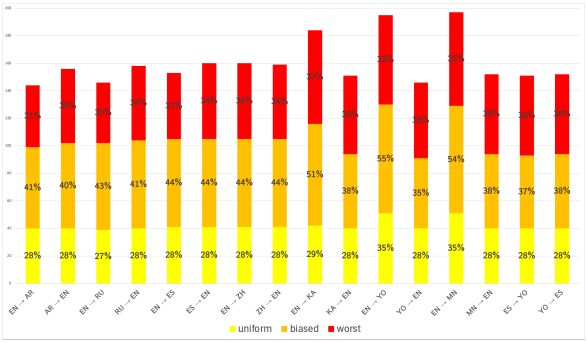


Figure 3: **Selection type distribution** This graph shows that the three EN→LRLs not only have more sentences, but also have way more *biased* sentences than other directions, which suggests a higher propensity to hallucinate.

## C Ablation study

The ablation study focus on hallucination **severity ranking**. We present results for comparability with Dale et al. (2023), which assesses the methods' abilities to accurately rank hallucinations by severity (e.g.,full hallucinations ranked higher than partial ones, and any hallucinations ranked above non-hallucinations). The employed metric is an adaptation of the ROC AUC for multiclass tasks, which calculates the percentage of incorrectly ranked sentence pairs with different labels and subtracts this value from the perfect score of 1. We compute these metrics separately for each translation direction to assess the detector's performance across different language pairs.

### C.1 LLMs for severity ranking

#### C.1.1 Prompt design

We designed tailored prompts for this approach, just as for our main binary approach, this time to generate multiclass predictions. For severity ranking, each prompt has a different assigned CoT.

**Severity ranking, Prompt1: *G-Eval* inspired** See Figure 4

**Severity ranking, Prompt2: *G-Eval* inspired, with *HalOmi*'s hallucination deifnition** See Figure 5

**Severity ranking, Prompt3: *G-Eval* inspired, with *HalOmi*'s hallucination deifnition, and language precision** See Figure 6

**Chain of Thoughts for severity ranking** See Figure 7 and Figure 8

*System:*
**Human Evaluation of Machine Translation Systems:**
**Hallucination Evaluation Criteria:** Does the translated text contain information unrelated to the source text?

- **No hallucination:** The translated text does not contain any hallucinated words.
- **Small hallucination:** The translated text contains 1-2 hallucinated words.
- **Partial hallucination:** The translated text includes at least 3 hallucinated words, but not all words are hallucinated.
- **Full hallucination:** Nearly all words in the translated text are hallucinated, with the exception of perhaps 1-2 words.

**Note:** The labels are mutually exclusive; for example, a translation with a partial hallucination does not qualify as a full hallucination.:
*User:*
Source Text: `src_text`
Translated Text: `mt_text`
Respond with the hallucination level label ONLY. Answer (Labels=['No hallucination','Small hallucination','Partial hallucination','Full hallucination'])

Figure 4: **Severity Ranking Prompt 1** - from *G-Eval*

#### C.1.2 Prompt evaluation

We evaluated three prompts and two CoT variations on the validation set to select the best prompt (Table 4). The prompt that achieved the highest average ROC AUC for both directions (DE↔EN) was chosen for each method. Subsequently, in the testing phase, each model was assessed with its optimal prompt.

### C.2 Embeddings for severity ranking

We computed the cosine similarity between the source text and machine-translated text embeddings for each embedding space and took the negative of these results. This approach ensures that hallucinations (indicative of embeddings that are farther apart) correspond to higher numbers, consistent with the ranking scale used in hallucination evaluation. Since this method does not require parameter tuning, the validation set was not utilized for thresholding in contrast to the binary approach.

### C.3 Results

In the same way as in the binary detection setting, the validation results Table 4 allowed to select the otpimal prompt for each LLM, and then evaluate this best prompt across the test set, using here the ROC AUC score. Testing results aredisplayed Table 5, and presents ROC AUC scores for all methods per translation direction. For HRLs, embeddings' high performance remains consistent with the binary hallucination approach. However, `BLASER-QE` remains the state-of-the-art in overall performance for severity ranking. The generalizability of these results requires further evaluation due to significant class imbalances in the dataset. Notably, in 11 of the 18 language directions, fewer than five samples are present in at least one hallucination severity category, see Appendix B.

## D Prompts

We used two types of CoTs: One based on the human guidelines for hallucination detection, and the other based on the severity level definition, that was readapted to each case. For **binary detection**, two CoTs were tested for three prompts.

**Binary detection, Prompt1 - from *G-Eval*** See Figure 9

**Binary detection, Prompt2 - from *G-Eval* with language precision** See Figure 10

**Binary detection, Prompt3 - Human designed prompt** See Figure 11

**Binary detection, Chain of Thoughts** See Figure 12 and Figure 13

## E LLMs experiments

### E.1 LLMs hyperparameters

For the evaluation of LLMs, we used *LangChain* to ensure reproducibility of results, except for `Llama3-70B` that was ran locally. We set the `TEMPERATURE` to `0` for minimum randomness and the `MAX_OUTPUT_TOKEN` to `15` to avoid verbose.All the experiments were zero-shot, with an exhaustive label (for example, *['Hallucination', 'No Hallucination']* for **binary detection**). These choices showed the highest performances in previous research (Kocmi and Federmann, 2023) (Wei et al., 2022).

| Model | *Prompt1* no CoT | *Prompt1* CoT1 | *Prompt2* no CoT | *Prompt2* CoT2 | *Prompt3* no CoT | *Prompt3* CoT2 | *AVG* Mean | *AVG* Std. |
|---|---|---|---|---|---|---|---|---|
| GPT4-Turbo | 0.78 | 0.70 | **0.83** | 0.81 | 0.82 | 0.81 | 0.79 | 0.05 |
| GPT4o | 0.81 | 0.82 | 0.83 | **0.83** | 0.83 | 0.83 | **0.83** | 0.01 |
| Command R | **0.82** | 0.79 | 0.77 | 0.80 | **0.83** | 0.79 | 0.80 | 0.02 |
| Command R+ | 0.75 | 0.76 | 0.79 | **0.80** | 0.77 | 0.75 | 0.77 | 0.02 |
| Mistral 8x22b | 0.57 | 0.58 | **0.77** | 0.67 | 0.69 | 0.67 | 0.66 | 0.07 |
| Sonnet | **0.82** | **0.83** | 0.82 | 0.79 | 0.81 | **0.83** | 0.82 | 0.01 |
| Opus | 0.79 | 0.80 | 0.82 | 0.83 | **0.85** | 0.77 | 0.81 | 0.03 |
| Llama3-70B | 0.80 | 0.81 | **0.81** | 0.78 | 0.81 | 0.78 | 0.80 | 0.01 |

Table 4: Validation results for hallucination detection across prompt variations for severity ranking.

| Model | EN→HRL AR | EN→HRL RU | EN→HRL ES | EN→HRL ZH | HRL→EN AR | HRL→EN RU | HRL→EN ES | HRL→EN ZH | EN→LRL KA | EN→LRL YO | EN→LRL MN | LRL→EN KA | LRL→EN YO | LRL→EN MN | ES→YO | YO→ES | *AVG* HRL | *AVG* LRL | *AVG* Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT text-embedding-3-large | 0.89 | 0.82 | 0.84 | **0.92** | 0.91 | 0.94 | 0.87 | **0.87** | 0.71 | 0.7 | 0.54 | 0.56 | 0.68 | 0.6 | 0.62 | 0.51 | **0.88** | 0.62 | 0.75 |
| Cohere Embed v3 | 0.84 | 0.87 | 0.83 | 0.88 | 0.9 | **0.96** | **0.89** | 0.83 | 0.75 | 0.73 | 0.54 | 0.58 | 0.74 | 0.64 | 0.65 | 0.59 | 0.88 | **0.65** | 0.76 |
| Mistral-embed | **0.92** | 0.88 | 0.82 | 0.85 | 0.92 | 0.86 | 0.86 | 0.83 | 0.72 | 0.7 | 0.56 | 0.53 | 0.68 | 0.61 | 0.63 | 0.53 | 0.87 | 0.62 | 0.74 |
| SONAR | 0.89 | 0.79 | **0.85** | 0.77 | 0.93 | 0.93 | 0.85 | 0.87 | 0.81 | 0.8 | 0.69 | 0.73 | **0.79** | 0.73 | **0.69** | **0.62** | 0.86 | 0.73 | 0.8 |
| GPT4-Turbo | 0.8 | 0.72 | 0.65 | 0.8 | 0.86 | 0.91 | 0.86 | 0.79 | 0.61 | 0.57 | 0.26 | 0.47 | 0.43 | 0.31 | 0.38 | 0.4 | 0.8 | 0.43 | 0.61 |
| GPT4o | 0.71 | 0.74 | 0.65 | 0.8 | 0.86 | 0.86 | 0.74 | 0.8 | 0.64 | 0.58 | 0.3 | 0.47 | 0.59 | 0.4 | 0.45 | 0.41 | 0.77 | 0.48 | 0.63 |
| Command R | 0.56 | 0.88 | 0.61 | 0.83 | 0.86 | 0.84 | 0.77 | 0.68 | 0.47 | 0.51 | 0.19 | 0.16 | 0.19 | 0.33 | 0.37 | 0.3 | 0.75 | 0.32 | 0.53 |
| Command R+ | 0.59 | 0.56 | 0.65 | 0.7 | 0.91 | 0.91 | 0.76 | 0.74 | 0.34 | 0.39 | 0.04 | 0.41 | 0.43 | 0.26 | 0.15 | 0.4 | 0.73 | 0.3 | 0.51 |
| Mistral 8x22b | 0.25 | 0.59 | 0.53 | 0.67 | 0.84 | 0.94 | 0.74 | 0.77 | 0.51 | 0.4 | 0.08 | 0.46 | 0.52 | 0.5 | 0.33 | 0.46 | 0.67 | 0.41 | 0.54 |
| Sonnet | 0.7 | 0.75 | 0.61 | 0.8 | 0.84 | 0.89 | 0.7 | 0.69 | 0.64 | 0.62 | 0.41 | 0.56 | 0.58 | 0.55 | 0.53 | 0.47 | 0.75 | 0.55 | 0.65 |
| Opus | 0.6 | **0.91** | 0.69 | 0.83 | 0.88 | 0.9 | 0.83 | 0.76 | 0.66 | 0.54 | 0.2 | 0.49 | 0.7 | 0.53 | 0.33 | 0.49 | 0.8 | 0.49 | 0.65 |
| Llama3-70B | 0.6 | **0.91** | 0.69 | 0.83 | 0.88 | 0.9 | 0.83 | 0.76 | 0.66 | 0.54 | 0.2 | 0.49 | 0.7 | 0.53 | 0.33 | 0.49 | 0.8 | 0.49 | 0.65 |
| BLASER 2.0-QE | 0.9 | 0.89 | 0.85 | 0.78 | **0.94** | 0.92 | 0.87 | 0.87 | **0.81** | **0.83** | **0.79** | **0.73** | 0.78 | **0.8** | 0.68 | 0.58 | 0.88 | **0.75** | **0.81** |

Table 5: ROC-AUC results for severity hallucination ranking across HRL and LRL directions.
*Bold values indicate the best performing prompt per model.*

### E.2 LLMs selection

We selected the following models for our evaluation: `GPT4-turbo`, widely adopted in both academic research and industrial applications due to its robust performance and versatility; `GPT4o`, the latest GPT model, optimised for better human-computer interaction; `Command-R`, known for its large context window, well-suited for tasks that require extended language understanding and generation; `Command R+`, an enhanced version of `Command R`, demonstrating strong performance in multilingual tasks, achieving impressive BLEU scores in benchmark datasets such as FLoRES and WMT23; `Mistral 8x22b`, currently the most performant open model from Mistral, excelling in various language tasks; `Claude Sonnet`, showing strong capabilities in multilingual tasks, similar to `Command R+`; `Claude Opus`, known as the "most intelligent" Claude model, offering advanced language understanding and generation capabilities; and `LLama3-70B`, the most capable openly available LLM from Meta, evaluated in its 70B size for comprehensive performance analysis. These models were chosen based on their demonstrated performance in various benchmarks and their potential to handle a wide range of language tasks effectively.

*System:*
You will be given a 'Source Text' and its translation, the 'Translated Text'. Your task is to identify the hallucination level of the translated text.
**Definition of Hallucination:**
A word in the translated text is considered a hallucination if it introduces information that is completely unrelated to the source text.
**Hallucination labels:**

- **No hallucination:** The translated text does not contain any hallucinated words.
- **Small hallucination:** The translated text contains 1-2 hallucinated words.
- **Partial hallucination:** The translated text includes at least 3 hallucinated words, but not all words are hallucinated.
- **Full hallucination:** Nearly all words in the translated text are hallucinated, with the exception of perhaps 1-2 words.

**Note:** The labels are mutually exclusive; for example, a translation with a partial hallucination does not qualify as a full hallucination.:
*User:*
Source Text: `src_text`
Translated Text: `mt_text`
Provide exactly one of the following hallucination level labels as your response. Do not include any additional text or explanation:

- No hallucination
- Small hallucination
- Partial hallucination
- Full hallucination

Figure 5: **Severity Ranking Prompt 2** - from *G-Eval* with Hallucination definition

*System:*
You will be given a 'Source Text' in `src_lang` and its translation in `tgt_lang`, the 'Translated Text'. Your task is to identify the hallucination level of the translated text. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
**Definition of Hallucination:**
A word in the translated text is considered a hallucination if it introduces information that is completely unrelated to the source text.
**Hallucination labels:**

- **No hallucination:** The translated text does not contain any hallucinated words.
- **Small hallucination:** The translated text contains 1-2 hallucinated words.
- **Partial hallucination:** The translated text includes at least 3 hallucinated words, but not all words are hallucinated.
- **Full hallucination:** Nearly all words in the translated text are hallucinated, with the exception of perhaps 1-2 words.

**Note:** The labels are mutually exclusive; for example, a translation with a partial hallucination does not qualify as a full hallucination.:
*User:*
Source Text: `src_text`
Translated Text: `mt_text`
Provide exactly one of the following hallucination level labels as your response. Do not include any additional text or explanation:

- No hallucination
- Small hallucination
- Partial hallucination
- Full hallucination

Figure 6: **Severity Ranking Prompt 3** - from *G-Eval* with Hallucination definition and language precision

**Evaluation Steps:**

1. Read the source text and the translated text carefully.
2. To decide whether the translated text contains hallucinations check if the source word “corresponds” to erroneous target tokens. For each work answer:
    - Does this source word fall into the common meaning category as this target word?
    - Does this source word have a semantic connection with this target word?
    - Can you try to come up with a reasonable theory on how this source word is associated with this target word?
    - If “no” to all the questions above, then hallucination. Keep a count of the number of hallucinated words for each sentence pair.
3. After reading all the source and translated text, assign a label to the pair based on the number of hallucinated words.

Figure 7: **Severity Ranking CoT 1** - from *HalOmi*'s human guidelines

**Evaluation Steps:**

1. Read the source text and the translated text carefully.
2. Initialize a counter ‘n = 0‘ for the number of hallucinated words.
3. For each word in the translated text, perform the following checks to determine if it is a hallucinated word:
    - Does this source word fall into the common meaning category as this target word?
    - Does this source word have a semantic connection with this target word?
    - Can you try to come up with a reasonable theory on how this source word is associated with this target word?
    - If "no" to all the questions above, then it is considered a hallucination. Increment ‘n‘ by 1.
4. After analyzing each word in the translated text:
    - If ‘n == 0‘, assign the label ’No hallucination’.
    - If ‘n‘ is 1 or 2, assign the label ’Small hallucination’.
    - If ‘n‘ is 3 or more but not all words are hallucinated, assign the label ’Partial hallucination’.
    - If nearly all words are hallucinated, assign the label ’Full hallucination’.’

Figure 8: **Severity Ranking CoT 2** - counting the number of hallucinated words

*System:*
Human Evaluation of Machine Translation Systems:
Hallucination Evaluation Criteria: Does the translated text contain information completely unrelated to the source text?
- *Hallucination*: there is hallucination.
- *No Hallucination*: there is no hallucination.

*User:*
Source Text: `src_text`
Translated Text: `mt_text`
Does the translation contain hallucination? Answer (label ONLY: 'Hallucination' OR 'No Hallucination'):

Figure 9: **Binary detection Prompt 1** - from *G-Eval*

*System:*
Instructions for Evaluating Machine Translation:
You will be given a source text in `src_lang` and a machine translated text in `tgt_lang`.
Your task is to identify if the machine translated text has hallucination or not.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Hallucination: Does the translated text contain information completely unrelated to the source text?
- *Hallucination*: there is hallucination.
- *No Hallucination*: there is no hallucination.

*User:*
Source Text: `src_text`
Translated Text: `mt_text`
Does the translation contain hallucination? Answer (label ONLY: 'Hallucination' OR 'No Hallucination'):

Figure 10: **Binary detection Prompt 2** - from *G-Eval* with language precision

*System:*
Instructions for Evaluating Machine Translation:
You will be given a source text in `src_lang` and a machine translated text in `mt_lang`.
Your task is to identify if the machine translated text has hallucination or not. Your task is to identify if the translated text is a hallucination. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
**Definition of Hallucination:** The translated text is considered a hallucination if it introduces information that is completely unrelated to the source text.
**Hallucination labels:**

- **Hallucination:** there is hallucination.
- **No hallucination:** there is no hallucination.

*User:*
Source Text: `src_text`
Translated Text: `mt_text`
Provide exactly one of the following hallucination labels as your response. Do not include any additional text or explanation:

- *Hallucination*
- *No hallucination*:

Figure 11: **Binary detection Prompt 3** - Human designed prompt

**Evaluation Steps:**
1. Read the source text and the translated text carefully.
2. To decide whether the translated text contains hallucinations check if the source tokens "correspond" to erroneous target tokens. For each token answer:

- Does this source word fall into the common meaning category as this target word?
- Does this source word have a semantic connection with this target word?
- Can you try to come up with a reasonable theory on how this source word is associated with this target word?

3. If "no" to all the questions above, then hallucination

Figure 12: **Binary detection - CoT1:** from *HalOmi*'s human guidelines

**Evaluation Steps:**

1. Read the source text and the translated text carefully.
2. Initialize a counter ‘n = 0‘ for the number of hallucinated words.
3. To decide whether the translated text contains hallucinations check if the source tokens "correspond" to erroneous target tokens. For each token answer:
    - Does this source word fall into the common meaning category as this target word?
    - Does this source word have a semantic connection with this target word?
    - Can you try to come up with a reasonable theory on how this source word is associated with this target word?
    - If "no" to all the questions above, then hallucination
4. After analyzing each word in the translated text:
    - If ‘n == 0‘, assign the label ’No hallucination’.
    - If ‘n‘ is 1 or more, assign the label ’Hallucination’.”’

Figure 13: **Binary detection - CoT2:** from *HalOmi*'s human guidelines and counting strategy

# F Binary detection results

## F.1 Validation results

Table 6 provides MCC scores per LLM for each of the prompts and CoT variations evaluated on the validation set. The most robust LLMs across prompt variations in the validation set, specifically Sonnet, `GPT4o`, and `Llama3-70B`, exhibit superior performance across language resource settings in the test set. This suggests that extensive prompt engineering might not be required for these models in the current task, as the performance using the optimal prompt from the validation set aligns with high performance on the test set.

## F.2 Test results

Figure 14 displays the performances of evaluated methods on the test set grouped by translation directions and resource setting. The results indicate that the highest scores for HRLs are achieved in translations to English, whereas for LRLs, the highest scores are from translations originating in English or Spanish. Additionally, these findings underscore that no single model uniformly excels across all translation directions.

| Model | *Prompt1* no CoT | *Prompt1* CoT1 | *Prompt2* no CoT | *Prompt2* CoT1 | *Prompt3* no CoT | *Prompt3* CoT1 | *Prompt3* CoT2 | *AVG* Mean | *AVG* Std. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | **Binary Detection (MCC)** | | | | | |
| GPT4-Turbo | 0.53 | 0.55 | **0.55** | 0.50 | 0.45 | 0.51 | 0.47 | 0.51 | 0.04 |
| GPT4o | 0.44 | 0.44 | **0.51** | 0.45 | 0.44 | 0.47 | 0.48 | 0.46 | 0.03 |
| Command R | 0.43 | 0.37 | 0.54 | 0.47 | 0.51 | 0.53 | **0.55** | 0.49 | 0.07 |
| Command R+ | 0.72 | 0.72 | 0.57 | 0.69 | 0.54 | **0.72** | 0.64 | 0.66 | 0.08 |
| Mistral 8x22b | 0.51 | 0.57 | 0.52 | 0.61 | **0.69** | 0.65 | 0.69 | 0.61 | 0.07 |
| Sonnet | 0.67 | 0.68 | **0.69** | 0.68 | 0.68 | 0.69 | 0.68 | **0.68** | 0.01 |
| Opus | 0.57 | 0.50 | 0.53 | 0.56 | **0.73** | 0.64 | 0.59 | 0.59 | 0.08 |
| Llama3-70B | 0.74 | 0.76 | 0.74 | 0.72 | **0.81** | 0.79 | 0.79 | 0.76 | 0.03 |

Table 6: Validation results for binary hallucination detection across prompt variations. Bold values indicate the best performing prompt per model. In the case of ties, we favor shorter prompts without CoT.

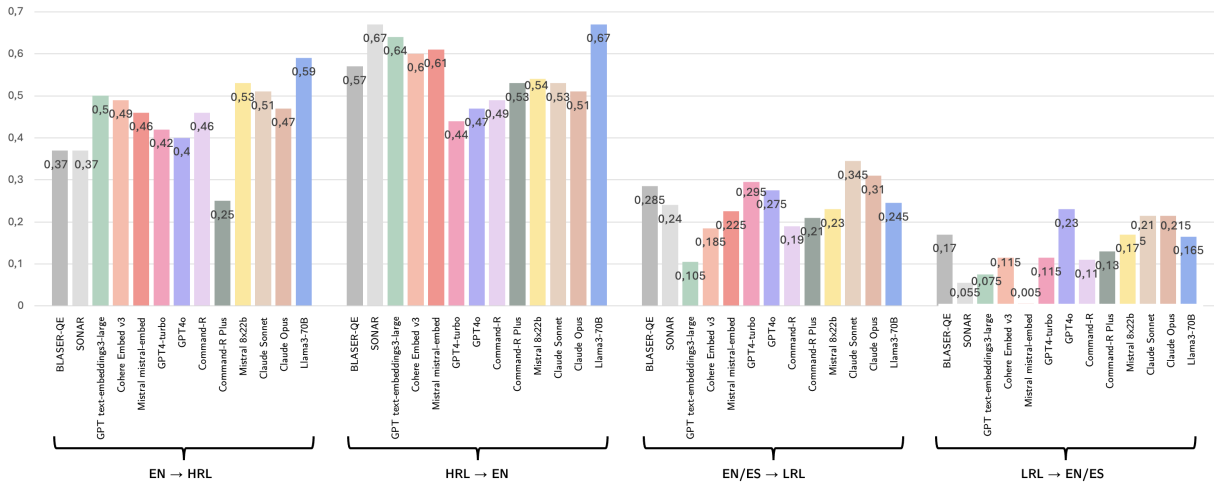


Figure 14: MCC average score across high and low resource levels, for different directions. The best performing models differ significantlly between HRLs and LRLs. For HRLs, `Llama3-70B` greatly outperforms other methods, whereas for LRLs, best performers differ from and to LRLs, with `Claude` and `GPT` models closely competing. Embeddings demonstrate impressive results, particularly for the `EN`→`HRL` directions.