HAT: Hallucination Annotation for Translation

Rajen Chatterjee, Xintong Li, Paisarn Charoenpornsawat, Allen Lee
Apple, Cupertino, CA USA
{rajen_c, xintongli, paisarn, alee87}@apple.com

Abstract

Hallucinations in machine translation (MT)—outputs that may be fluent yet unfaithful to the source content—remain a critical obstacle. They hinder the reliable deployment of MT systems in real-world applications. Despite growing attention to this phenomenon, progress has been constrained by the lack of large-scale, high-quality benchmarks dedicated to hallucination detection. We introduce **HAT** (**H**allucination **A**nnotation for **T**ranslation), a novel dataset designed to advance research on this problem. HAT comprises 350,959 span-level annotated samples across 38 language pairs, with approximately 8,000–10,000 samples per pair partitioned into training, development, and test sets. Annotations were produced by professional translators under rigorous quality control protocols to ensure reliability. We provide a detailed analysis of hallucination distributions and establish benchmark performance using a diverse set of baselines, including automatic MT evaluation metrics as well as large language models. By providing the first large-scale, systematically annotated resource for hallucination detection in MT, HAT enables the development of more faithful translation models and lays the groundwork for future research on building trustworthy machine translation systems.

1 Introduction

Machine translation (MT) has seen dramatic progress in recent years, with large language models (LLMs) achieving near state-of-the-art performance across multiple languages (Kocmi et al., 2024). Nevertheless, MT systems—whether encoder–decoder or decoder-only architectures—remain prone to hallucination, where the model produces fluent but fabricated content that has little or no connection to the source sentence. Although hallucinations are relatively rare, their occurrence can undermine user trust and raise serious safety concerns.

Hallucination in MT has long been a subject of study, with prior work focusing on probing models for hallucination, designing hallucination detection models, and developing mitigation strategies (Lee et al., 2019; Xu et al., 2023; Sennrich et al., 2024; Dale et al., 2023a; Guerreiro et al., 2023a,c; Raunak et al., 2021; Wang and Sennrich, 2020; Dale et al., 2023b; Himmi et al., 2024; Benkirane et al., 2024; Tang et al., 2025; Waldendorf et al., 2024). Central to all these efforts is a hallucination detection metric, which underpins the reliability of conclusions drawn from such studies. Building a robust detection model, however, requires access to hallucination examples for training or fine-tuning, as well as a gold-standard evaluation set for benchmarking. Since naturally occurring hallucinations are rare, collecting them at scale across multiple languages is resource- and compute-intensive. An alternative approach is to introduce synthetic noise into the input to trigger hallucinations, but it remains unclear whether synthetic artifacts adequately represent real-world cases, limiting the generalizability of such datasets.

As a result, most prior work has relied on proxy metrics: general MT quality metrics (e.g., BLEU Papineni et al., 2002, chrF Popović, 2015, COMET QE Rei et al., 2020), sentence similarity metrics (e.g., LASER3 Heffernan et al., 2022, LaBSE Feng et al., 2022, BLASER 2.0 QE Chen et al., 2023), intrinsic MT model signals (e.g., sequence log probability, ALTI+ Ferrando et al., 2022, Attn-OT Guerreiro et al., 2023b), or judgments from LLMs (e.g., GPT4, LLaMA3) (Benkirane et al., 2024). While useful, these proxies are not explicitly optimized for hallucination detection.

A small number of resources with naturally occurring hallucination annotations do exist; however, they focus on only one language pair or the volume of data is too low. For example, Guerreiro et al., 2023c released 3,015 annotated samples for German→English, covering categories such as

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15865–15888

mistranslations (e.g., named entities, omissions) and hallucinations (e.g., oscillation, strongly detached, fully detached). Zhou et al., 2021 released Chinese→English data with token-level hallucination labels. Similarly, Dale et al., 2023b released HalOmi dataset which includes 2,866 samples across 18 language pairs, annotated for hallucination and omission. However, this dataset provides around 150 samples per pair, limiting its utility for developing strong hallucination detection model.

To address this gap, we introduce HAT (Hallucination Annotation for Translation), a large-scale multilingual benchmark with 350,959 samples annotated for hallucination at span-level across 38 language pairs—roughly 8,000–10,000 per pair—partitioned into training, development, and test sets. By significantly increasing both language coverage and dataset size, HAT provides the scale necessary for building robust hallucination detection models. Beyond releasing this resource, we also benchmark a range of automatic MT evaluation metrics and LLMs on the task, establishing strong baselines for future research.

2 Data Creation

Since our goal is to build a large-scale benchmark, we focus exclusively on the hallucination category, rather than including other error types such as omissions or mistranslations. Narrowing the scope to a single phenomenon enables annotators to concentrate more effectively, leading to higher-quality annotations, while also keeping the task design straightforward and budget-feasible.

The creation of this resource was guided by the following requirements:

- **Broad language coverage** – include a diverse set of language pairs to obtain a holistic view of hallucinations.
- **Sufficient scale** – produce enough annotated samples to support both training/fine-tuning and evaluation of detection models.
- **Natural hallucinations** – collect naturally occurring hallucinations rather than relying on artificially perturbed inputs.
- **High-quality MT systems** – use strong translation systems across all languages to minimize noise from low-quality baselines.
- **Expert annotation** – rely on professional translators and enforce rigorous quality-control protocols to ensure reliable labels.

Guided by these requirements, we selected 20 languages, for which we maintain strong in-house MT systems. For each language, we constructed datasets in both English→X and X→English directions, resulting in a total of 38 language pairs. The languages include Arabic, Dutch, English, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Simplified Chinese, Spanish, Thai, Traditional Chinese, Turkish, Ukrainian, Vietnamese.

2.1 Data Creation and Translation

For each language, we crawled approximately 5 million sentences from the web. The data was then deduplicated, passed through a language-identification filter (Joulin et al., 2017a,b), and further cleaned with length-based filtering to remove sentences that were too short or excessively long. The cleaned source sentences were translated into the target language using an in-house encoder–decoder neural MT system. We opted for this system over LLM for two reasons: (i) it delivers strong performance across all languages; and (ii) it provides significantly faster inference, which was essential given the need to translate more than 100 million sentences across all pairs.

After cleaning and translation, we applied data sampling heuristics (described in Section 2.2) to select $10,000$ candidate samples per language pair. These were submitted for human annotation following the procedures outlined in Section 3.

2.2 Data Sampling

Manually annotating millions of translations is infeasible, and given that hallucinations are a rare phenomenon, random sampling would yield very few hallucinated examples. To address this, and following strategies from prior work, we score translations using multiple hallucination detection models and then apply a sampling heuristic to increase the likelihood of including hallucinated translations in the dataset submitted for annotation.

Specifically, we employ the following reference-free models: [1]

[1]Models were selected based on their permissive licensing, which also facilitates unbiased benchmarking of more advanced models.

- COMET-QE (Rei et al., 2021): a translation quality estimation metric based on a cross-lingual encoder, trained with a regressor to minimize mean squared error.[2]

- LaBSE (Feng et al., 2022): a cosine similarity metric using a dual-encoder transformer architecture with an additive margin softmax loss.[3]

- LASER2 (Heffernan et al., 2022): a cosine similarity metric based on an LSTM encoder–decoder architecture.[4]

- ALTI+ (Ferrando et al., 2022): an input token attribution method that evaluates token contribution to the generated translation using attention scores from the same MT model that produced the translation.[5]

- TopNGram (Raunak et al., 2021): a binary classifier that outputs 1 if the count of the most repeated n-gram in the translation exceeds the top repeated n-gram count in the source by at least a threshold t (we set $n = 3$ and $t = 2$).

- MT Sequence Log Probability (Guerreiro et al., 2023c): length-normalized sequence log probability from the generating MT model.

For sampling, we adopt a heuristic designed to select up to $10,000$ samples per language pair without replacement:

- For each model, select $1,000$ samples with the lowest quality scores.

- For each model, select $1,000$ additional samples randomly from the lowest quality 1st percentile of scores, excluding those already selected in step 1.

- If the total number of selected samples exceeds $10,000$ after step 2, we randomly downsample to $10,000$ to obtain the final set of samples, which is partitioned into $5,000$, $2,000$, and $3,000$ samples respectively in training, development, and test sets.

This sampling approach prioritizes examples that are more likely to contain hallucinations while maintaining diversity across models and translation characteristics.

3 Annotation

3.1 Task Requirements

To ensure the reliability of our hallucination annotations, we established strict requirements covering annotator profile, vendor responsibilities, and platform setup. Vendors were required to demonstrate global reach and the ability to source qualified annotators across all languages. All annotators were required to be native speakers of the target language and fluent in the source language. For rare or low-resource language pairs, professional linguists were engaged to guarantee accuracy. The task interface displayed localized instructions and examples for each language pair, allowing annotators to reference guidelines while working. Vendors were also responsible for project management, maintaining direct communication with our team, and ensuring annotation quality consistently exceeded 95% quality metric. The annotation process was conducted on a secure, vendor-managed platform, with no reliance on third-party providers for crowdsourcing or hosting. The platform supported:

- API integration for automated data upload and retrieval,

- project and annotator tracking,

- double-review workflows, where a second reviewer could automatically validate annotations,

- queue management for rejected samples, which were re-annotated at no additional cost,

- customizable task UIs, aligned with provided mock-ups and localized guidelines.

This integrated set of requirements ensured that the annotation process was both linguistically rigorous and operationally robust, providing a solid foundation for dataset creation.

3.2 Quality Control

A comprehensive quality control pipeline was implemented to guarantee annotation consistency and accuracy. Before beginning the task, annotators were required to pass a 25-sample exam covering a range of hallucination severities in translation:

[2] https://huggingface.co/zwhe99/wmt21-comet-qe-da

[3] https://huggingface.co/sentence-transformers/LaBSE

[4] https://github.com/facebookresearch/LASER

[5] https://github.com/mt-upc/transformer-contributions-nmt

- 2 samples with 90–100% of hallucination text,
- 10 samples with 40–60% of hallucination text,
- 10 samples with 10–20% of hallucination text,
- 2 samples with mistranslation,
- 1 sample of a perfect translation.

Six of the exam samples required multiple hallucination spans to be identified (4 from the 40–60% group and 2 from the 10–20% group). Annotators were evaluated on precision and recall metrics, both of which had to meet or exceed 95%. Failing the exam was permitted only once; a second failure resulted in disqualification.

Annotation proceeded in three stages:

1. Initial annotation – annotators labeled hallucination spans in the translations (one annotation per sample).
2. Verification round – a second reviewer checked annotations, corrected minor errors, or sent problematic samples back for re-annotation.
3. Final validation – random subsets of annotated data were re-checked, and precision/recall scores were recomputed. If the batch failed to meet the 95% threshold, the entire batch was re-annotated.

3.3 Annotation Schema

Each annotated sample followed a structured schema with the following key fields:

- srcLocale / tgtLocale: the source and target locale code.
- id: unique sample identifier.
- annotatorID: anonymized annotator identifier.
- source / target: raw source and translated texts.
- errors: structured error entries, defined by:
  - category: error category (set to "hallucination" in this task).
  - spans: list of hallucinated spans, each with a start index, end index, and span text.

Additional metadata included source-level error labels (e.g., wrong language, nonsensical input, spelling errors, mixed languages) and skip reasons (e.g., sensitive content, annotator uncertainty, highly specialized text). These fields allowed problematic cases to be flagged during annotation.

3.4 Guidelines

The hallucination annotation guidelines are motivated from prior work (Dale et al., 2023b). We define hallucinations as any translation content that is unrelated to the source text, distinguishing them from mistranslations, which still retain some semantic link to the source. Annotators are advised to follow a conservative rule: if uncertain, don't label as hallucination. To guide this distinction, the guidelines propose a set of decision questions:

- Does any source text span belong to the common meaning category of the translation span?
- Does any source text span have a semantic connection with this translation span irrespective of the context?
- Can you try to come up with a reasonable theory on how any source text span can be associated with this translation span?

If the answer is "no" to all, the translation span is considered a hallucination.

The guidelines also included localized examples for each language pair to illustrate the difference between mistranslations and hallucinations. In Spanish–English, "*Este es el árbol más alto del bosque*" ("*This is the tallest tree in the forest*") translated as "*This is a big tree in the forest*" is a mistranslation, since "*big*" and "*tall*" share meaning, whereas "*This is the tallest and oldest tree in the forest*" is a hallucination, with "*and oldest*" adding ungrounded content. Similarly, "*La aplicación falló debido a un error*" ("*The application crashed due to a bug*") translated as "*The application crashed due to an insect*" is a mistranslation, but "*The application on my phone crashed due to a bug*" is a hallucination, as "*in my phone*" introduces unrelated information. More examples are provided in Appendix Table 7. These guidelines, reinforced by examples across multiple languages, ensure annotators apply consistent reasoning when identifying hallucinations.

The annotation process is span-based, requiring annotators to mark hallucinated content with precise start and end boundaries. Continuous hallucinations must be captured as a single span, while

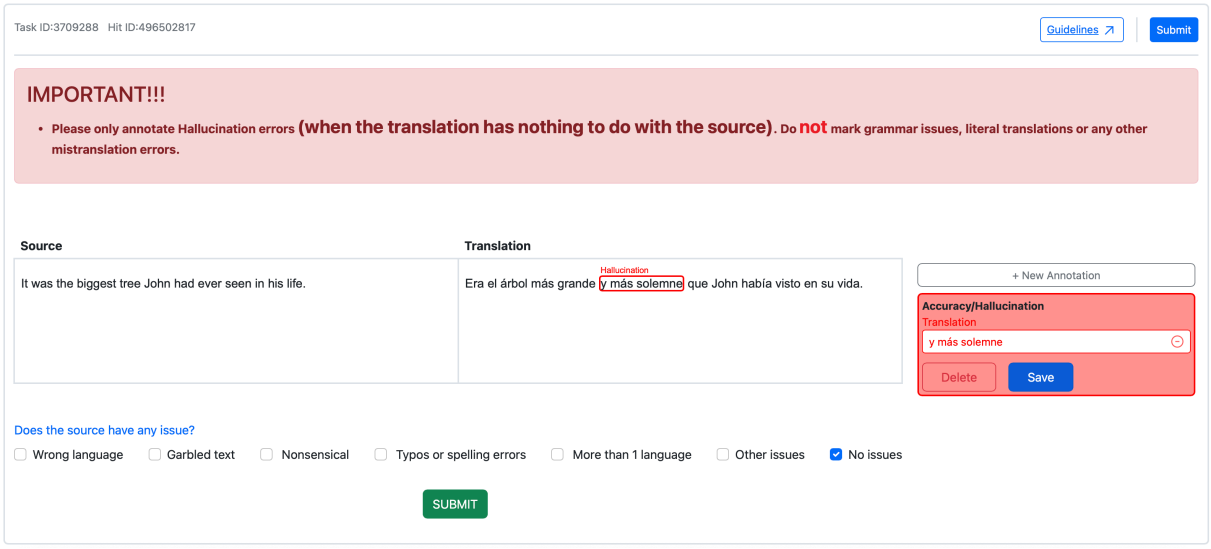
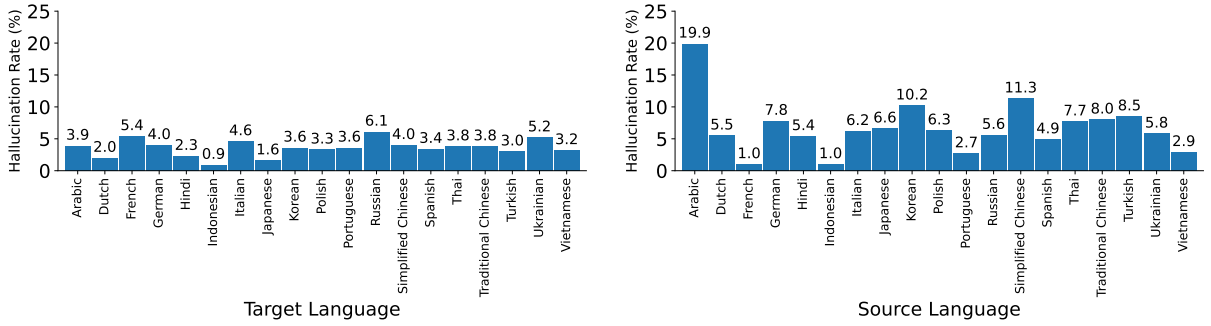


Figure 1: HAT Annotation UI



(a) English→X

(b) X→English

Figure 2: Sentence-level hallucination rate (%) for the HAT test set

discontinuous hallucinations are annotated using multiple spans. The task includes only one error category—hallucination—but annotators are also instructed to flag special source-side issues such as wrong language, garbled or nonsensical text, typos, and mixed-language content. Sentences with issues on source side are discarded from final dataset.

Figure 1 presents the annotation user interface (UI), which is divided into three main sections. The first section provides general instructions along with a link to the full annotation guidelines. The second section displays the actual task, showing the source sentence alongside its corresponding translation. If hallucinations are present in the translation, the annotator highlights the relevant span (illustrated with a red box in the figure). Each selected span then appears in the right-hand panel, where the annotator can choose to save or delete the annotation. The third section records any source-side issues; if none are present, the annotator must explicitly select the "No issues" option.

The sentence-level hallucination rate (%) for the HAT test set is shown in Figure 2a and Figure 2b, corresponding to English→X and X→English directions, respectively. The rate is calculated as the proportion of test samples flagged as containing hallucinations relative to the total number of samples. Overall, we find that hallucination rates are higher when English is the target language. A likely explanation is that the hallucination detector models used during the data sampling stage (see Section 2.2) are more effective at identifying hallucinations when the target language is English. Additional statistics are provided in Appendix A.1, including final number of samples retained after the annotation process, character-level hallucination rates, source-to-target length ratios, and sentence-level hallucination distributions across different hallucination score ranges.

4 Benchmark

We establish a sentence-level hallucination detection benchmark on the HAT test set to evaluate and compare the performance of different systems across all 38 language pairs. Specifically, we assess both automatic MT evaluation models and LLMs, including open- and closed-source variants. To ensure a comprehensive and unbiased analysis, we employ multiple metrics. In the HAT datasets, sentence-level labels are binary: 0 indicates no hallucination, while 1 indicates the presence of hallucination.

4.1 Metrics

We consider the following metrics:

- **Precision**: Proportion of predicted positives that are truly positive, measuring exactness.
- **Recall**: Proportion of actual positives correctly identified, measuring completeness.
- **F1 Score:** Harmonic mean of precision and recall, balancing the ability to identify positives and ensure predicted positives are correct. We use binary average that reports score for positive class i.e. hallucination.
- **Matthews Correlation Coefficient (MCC):** Correlation-based metric considering all confusion matrix entries, providing a balanced score even under severe class imbalance.

4.2 Models

We evaluate two groups of systems for hallucination detection task:

- **Neural models:** These systems produce continuous scores, which we convert into binary predictions using thresholds tuned on the development set for each model–metric pair.[6]
  - BLASER-2.0-QE (Chen et al., 2023): A sentence similarity model based on SONAR embedding model (Duquenne et al., 2023) which has been shown to perform state-of-the-art on hallucination detection task (Dale et al., 2023b).[7]
  - METRICX-24-HYBRID-XXL-QE (Freitag et al., 2024): A state-of-the-art MT evaluation model based on mT5-XXL pre-trained language model (Xue et al., 2021).[8]
  - XCOMET-XXL-QE (Guerreiro et al., 2024): Another state-of-the-art MT evaluation model based on XLM-R-XXL pre-trained language model (Conneau et al., 2020).[9]
- **Large language models:** We evaluate several LLMs of varying sizes, including GPT-OSS-120B, GPT-OSS-20B, Qwen3-235B-A22B, Qwen3-8B, DeepSeek-V3.1, Gemini-2.5-Flash, and Gemini-2.5-Pro. A single, consistent prompt derived from our annotation guidelines is used across all models. The prompt specifies the task definition, key aspects to consider, the distinction between hallucination and mistranslation, and the expected response format. It also includes illustrative examples in English–Spanish: five highlighting hallucinations and five showing mistranslations that should not be flagged as hallucinations. The full prompt is provided in the Appendix (Figure 8). We also compare this prompt with a simple baseline prompt as discussed in Section A.2. For inference, we use the same hyperparameters across all models: temperature $= 0$ and top-$p = 1$.

4.3 Results and Analysis

4.3.1 Holistic View

In Figure 3, we present the benchmark performance of hallucination detection models across multiple metrics, averaged over all language pairs. LLMs consistently outperform neural baselines on most metrics, with the exception of precision, where some LLMs perform comparably to neural models. These results establish LLMs as the new state of the art for hallucination detection. Among them, Gemini-2.5-Pro achieves the strongest results, followed by Gemini-2.5-Flash, as reflected in F1 and MCC scores. Across all metrics, the best LLM outperforms the strongest neural baseline by up to 2.5×. Most models perform better on recall than on precision; notably, Gemini-2.5-Pro exhibits the lowest recall among LLMs, likely reflecting a trade-off for higher precision. Among open-source LLMs, GPT-OSS-120B with 5B active parameters

[6] For precision and recall, we use the same threshold as F1.

[7] https://huggingface.co/facebook/blaser-2.0-qe

[8] https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6-bfloat16

[9] https://huggingface.co/Unbabel/XCOMET-XXL

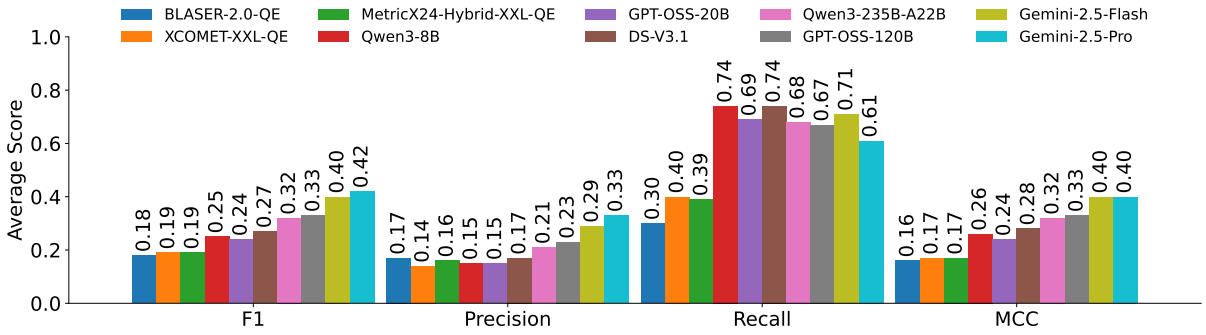


Figure 3: HAT benchmark results on test sets (averaged over all language pairs).

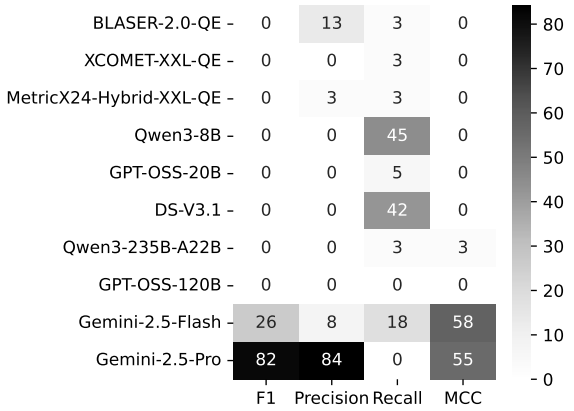


Figure 4: Win rate (%) over all language pairs.

surprisingly matches or outperforms both Qwen3-235B-A22B and DeepSeek-V3.1, which have 22B and 37B active parameters, respectively. Among neural models, BLASER-2.0 was previously considered state of the art, but more recent metrics such as XCOMET-XXL and MetricX-24-Hybrid now match or surpass its performance on most metrics.

Models often excel on different language pairs, with one model leading on some pairs while another performs best on others. This makes cross-lingual comparison non-trivial. To address this, we compute the *win rate* for each metric. For every language pair and metric, we identify the model(s) achieving the top score. A model's win rate is defined as the number of language pairs in which it ranks highest, divided by the total number of pairs. Ties are included, with all models sharing the top position receiving credit.[10] Figure 4 presents win rates (%) as a heatmap. Gemini-2.5-Pro dominates overall, with win rates of 82%, 84%, and 55% on F1, precision, and MCC, respectively, followed by Gemini-2.5-Flash. For recall, performance is more evenly distributed: Qwen3-8B ranks first with 45%, followed by DS-V3.1 with 42%.

Comprehensive results for all models and metrics across all language pairs are provided in the Appendix: Table 3, Table 4, Table 5, and Table 6 report performance on F1, precision, recall, and MCC, respectively.

4.3.2 Language Group Analysis

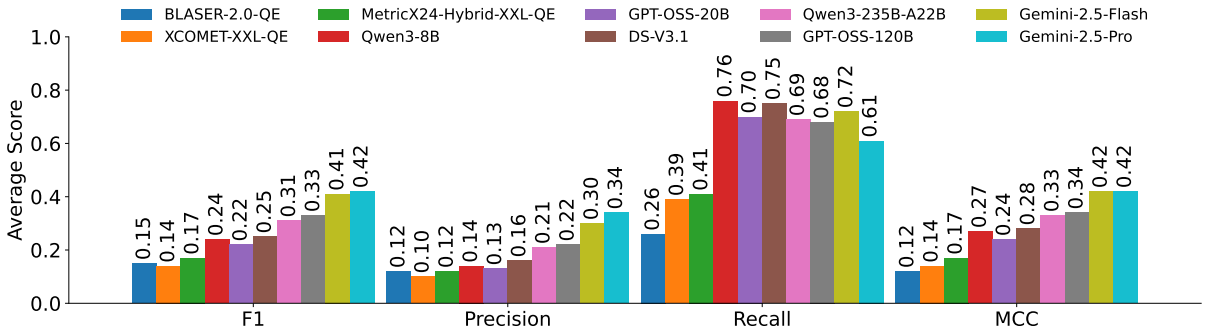
To better understand regional variation in model performance, we divide the evaluation into two script-based groups, with English included in both:

- **Latin:** Dutch, French, German, Indonesian, Italian, Polish, Portuguese, Spanish, Turkish, Vietnamese
- **Non-Latin:** Arabic, Hindi, Japanese, Korean, Russian, Simplified Chinese, Thai, Traditional Chinese, Ukrainian

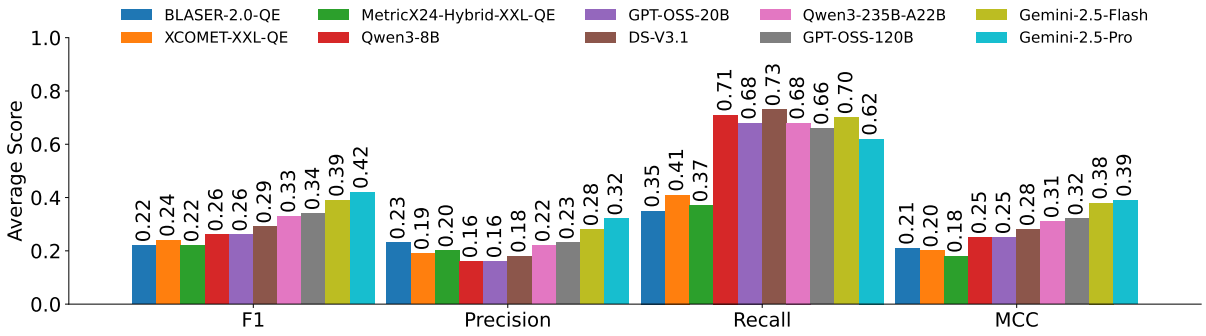
Figure 5 shows benchmark results (F1, precision, recall, and MCC) averaged within each group. Overall, model rankings remain consistent across both groups: Gemini-2.5-Pro and Gemini-2.5-Flash lead in both groups, with the former achieving the highest scores. For neural metrics, performance is relatively weaker in the Latin group. This effect is largely driven by language pairs with very low sentence-level hallucination rates (around 1%), such as French→English, Indonesian→English, and English→Indonesian, which lead to poor precision scores. While these pairs also challenge LLMs, LLM-based models still perform more robustly across both groups.

To further examine group-specific behavior, Figure 6 reports win rates (%) of each model. Gemini-2.5-Pro again leads across most metrics, reaching an impressive 75% and 89% win rate in F1

[10] As ties are included, win rates within a group may sum to more than 100%.

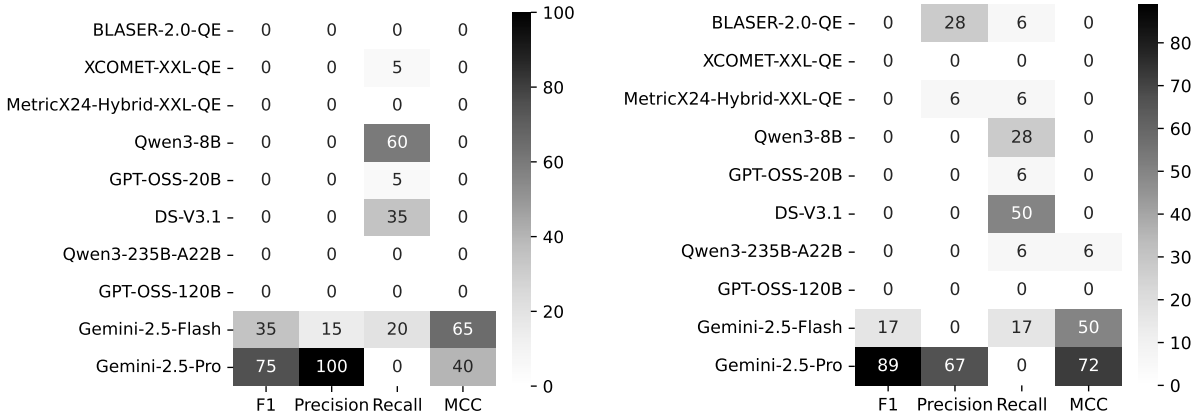


(a) Latin language group



(b) Non-Latin language group

Figure 5: HAT benchmark results on test sets (averaged over language pairs within each language group).



(a) Latin language group

(b) Non-Latin language group

Figure 6: Win rate (%) of models within each language group.

metric within the Latin and Non-Latin group, respectively. Surprisingly, Gemini-2.5-Flash outperforms Gemini-2.5-Pro in MCC metric within the Latin group, achieving a win rate of 65%, while Gemini-2.5-Pro maintains a strong lead in the Non-Latin group with a win rate of 72%. Finally, traditional neural metrics remain competitive in specific contexts. BLASER-2.0-QE, for example, shows strong precision in the Non-Latin group achieving second rank after Gemini-2.5-Pro, particularly for Japanese, Korean, and Traditional Chinese, highlighting the continued relevance of specialized baselines in certain linguistic settings.

4.4 Error Analysis

To better understand model errors, we analyzed the outputs of Gemini-2.5-Pro on the Spanish→English language pair, as this model achieved the highest overall performance.

At a high level, most examples incorrectly flagged as hallucinations (false positives) were actually mistranslations or minor textual mismatches

rather than true hallucinations. For instance, the Spanish sentence *"Ayudas y subvenciones de Artes escénicas y música"* paired with *"Grants and grants for Performing Arts and Music"* was flagged due to the repeated word "Grants," even though the repetition reasonably corresponds to the source terms "Ayudas" and "subvenciones." Similarly, *"Platos Fuertes con Acompañamientos • Main Courses with Sides"* paired with *"Main Courses with Sides • Main Courses with Sides"* was flagged because of repeated phrasing, despite being a faithful translation. Other false positives reflected minor mistranslations, such as *"YOURS Sudadera gris jaspeado con cuello vuelto"* paired with *"YOURS Taylor gray sweatshirt with a turned neck"*, where the color "Taylor" incorrectly replaced "Heather."

False negatives—actual hallucinations missed by the model—were often perceived as mistranslations, as indicated by the reason field in the LLM output. For example, *"Encontrarás botas de agua baratas"* paired with *"You will find cheap weer boots"* was marked as a mistranslation, although the word "weer" is a clear hallucination. Similarly, *"Circulando se ven las tacomas"* paired with *"You can see the tacos circulating"* was considered a mistranslation because "tacomas" (a truck brand) was interpreted as "tacos" (a food item). Other missed hallucinations include *"Estuve trabajando en mili mall"* paired with *"I was working at the military mall"* (where "military" is hallucinated for "mili"), and *"Barrabás tiene esclavos y charreteras"* paired with *"Barrabbas has slaves and troopers"* (where "troopers" is hallucinated instead of "epaulettes").

These examples highlight two key challenges for hallucination detection: distinguishing repeated or slightly altered phrases from hallucinations, and correctly identifying subtle lexical substitutions that constitute true hallucinations.

5 Conclusion

We presented HAT, the first large-scale, multilingual, span-level annotated dataset for hallucination detection in machine translation. HAT contains 350,959 professionally annotated samples across 38 language pairs, with roughly 8,000–10,000 samples per pair, and is partitioned into training, development, and test sets. The dataset benefits from rigorous quality control protocols, expert translator annotation, and carefully curated guidelines that capture precise span-level hallucinations. Its scale, linguistic diversity, and high-quality annotations make HAT a robust foundation for both training hallucination detection models and benchmarking across a broad range of languages.

Using HAT, we establish a comprehensive benchmark for sentence-level hallucination detection. LLMs consistently outperform neural metrics on F1 and MCC, with Gemini-2.5-Pro leading overall, closely followed by Gemini-2.5-Flash. Recall exhibits more variability, with Qwen3-8B and DeepSeek-V3.1 excelling in specific language groups. Among open-source LLMs, GPT-OSS-120B performs remarkably well, matching or surpassing larger models like Qwen3-235B-A22B and DeepSeek-V3.1. Among neural models, recent MT evaluation models like XCOMET-XXL and MetricX-24-Hybrid slightly outperform the previously dominant BLASER-2.0-QE overall. However, BLASER-2.0-QE remains competitive, achieving the highest precision in the Non-Latin group.

HAT not only sets a new standard for dataset quality and multilingual coverage but also motivates further research into hallucination detection and mitigation. By providing a reliable benchmark, it enables the development of models that produce faithful and trustworthy translations, improving safety, user trust, and downstream application reliability. Additionally, HAT lays the groundwork for future extensions, including richer error annotations, severity levels, span-level hallucination modeling, and integration into real-world MT workflows. We believe this resource will serve as a catalyst for the community, accelerating progress toward translation systems that are both fluent and accurate.

6 Limitations

While HAT represents a substantial advance, it also has a few limitations that point to promising directions for future work.

- First, the dataset primarily covers high- and medium-resource languages, leaving low-resource languages underrepresented. Expanding coverage in this area would further enhance the benchmark's inclusivity and impact.

- Second, our benchmark focuses on sentence-level hallucination detection. Extending it to span-level evaluation would provide a more fine-grained understanding of hallucinations

and better reflect real-world translation challenges.

- Third, our exploration of LLM prompting strategies was limited to a couple of prompts. More systematic prompt engineering, including localized few-shot examples, may yield improved performance and insights.

- Finally, all models in this study were evaluated in their off-the-shelf form, without additional training or fine-tuning. Since HAT includes training and development splits, future work can investigate how model adaptation specifically for hallucination detection affects performance.

7 Ethical considerations

This work promotes safer MT by releasing a publicly available benchmark for hallucination detection. Data was drawn from public sources and professionally annotated, with precautions taken to avoid sensitive information. Nonetheless, the dataset may underrepresent low-resource languages, reflect cultural or linguistic biases, and contain artifacts from web or model-generated text. We encourage responsible use of HAT and caution against deploying models trained on it without further validation.

Acknowledgments

We thank Liling Tan, Stephan Peitz, and Stephen Pulman for their thorough internal review and valuable feedback. We are also grateful to the entire machine translation team at Apple for developing the frameworks that enabled our large-scale experiments. Finally, we sincerely thank Matthias Paulik for sponsoring this work, which supported the creation of the annotations.

References

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665, Miami, Florida, USA. Association for Computational Linguistics.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

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

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint*.

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

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

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami,

Florida, USA. Association for Computational Linguistics.

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

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

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023c. 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.

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

Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno M Guerreiro. 2024. Enhanced hallucination detection in neural machine translation through simple detector aggregation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18573–18583, Miami, Florida, USA. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jegou, and Tomas Mikolov. 2017a. Fasttext.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017b. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Zilu Tang, Rajen Chatterjee, and Sarthak Garg. 2025. Mitigating hallucinated translations in large language

models with hallucination-focused preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3410–3433, Albuquerque, New Mexico. Association for Computational Linguistics.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, St. Julian's, Malta. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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

A Appendix

A.1 HAT Dataset Statistics

Tables 1 and 2 report dataset statistics for all language pairs and splits, covering English→X and X→English directions, respectively. We use the following acronyms for column names:

- **NS**: Number of samples in the split.
- **T2SLR**: Target-to-source length ratio (in characters). Values greater than 1 indicate that the target text is longer than the source.
- **T2SLRHG**: Target-to-source length ratio for the hallucination subset. Values greater than 1 indicate that the hallucinated target text is longer than the source.
- **CHR**: Character-level hallucination rate — proportion of characters flagged as hallucination across all translations in a given split.
- **SHR**: Sentence-level hallucination rate — proportion of all samples in a given split that contain hallucinations.
- **SHR (ID1–ID2]**: Sentence-level hallucination rate for specific hallucination score ranges — proportion of samples whose hallucination scores fall within the range (ID1, ID2]. Each sample's *hallucination score* is defined as the number of characters flagged as hallucinated divided by the total number of characters in the translation. The percentage is calculated relative to the total number of hallucinated samples in a given split.

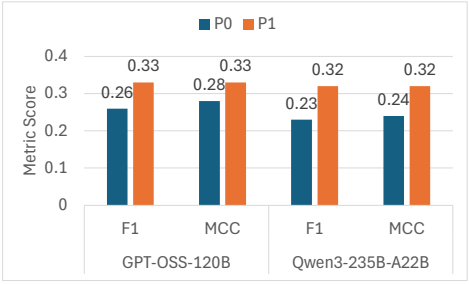


Figure 7: Prompt comparison.

A.2 LLM Prompt Comparison

We investigate the role of prompt design by comparing a simple baseline prompt (P0) shown in Figure 9 with our guideline-based prompt (P1). Prompt P0 contains only a generic instruction on hallucination detection, without explicitly distinguishing hallucinations from mistranslations and without examples. In contrast, P1 integrates structured guidelines and illustrative examples (as discussed in Section 4.2).

We evaluate GPT-OSS-120B and Qwen3-235B-A22B under both prompts using F1 and MCC, averaged across all language pairs (Figure 7). Results show that P1 consistently and substantially improves performance for both models. F1 score of GPT-OSS-120B and Qwen3-235B-A22B relatively improves by 27% and 39%. Interestingly, under P0, GPT-OSS-120B leads by 0.03 F1 points, but when switching to P1, the performance gap reduces to 0.01 F1 points. The greater performance

gain for Qwen3-235B-A22B indicates that its architecture is more sensitive to, and benefits more from, structured prompt design.

These findings highlight two important implications. First, careful prompt engineering is not only beneficial but sometimes necessary for accurate hallucination detection. Second, prompt quality can shift the relative ranking of models, suggesting that model comparisons without standardized, well-designed prompts may underestimate the true capabilities of certain systems.

Table 1: HAT dataset statistics (English→X).

Target Language	Split	NS	T2SLR	T2SLRHG	CHR (%)	SHR (%)	SHR (%) (0-5]	SHR (%) (5-25]	SHR (%) (25-100]
Arabic	dev	1816	0.9	1.4	5.6	4.5	9.9	35.8	54.3
	test	2733	0.9	1.1	2.6	3.9	5.6	30.8	63.6
	train	4541	0.9	1.4	4.7	3.4	5.8	36.5	57.7
Dutch	dev	1862	1.1	1.2	0.6	1.4	15.4	23.1	61.5
	test	2774	1.1	1.2	0.6	2.0	14.5	27.3	58.2
	train	4615	1.1	1.2	0.5	1.9	12.5	22.7	64.8
French	dev	1853	1.2	1.1	0.7	2.9	5.6	38.9	55.6
	test	2773	1.2	1.4	1.6	5.4	6.6	33.1	60.3
	train	4624	1.2	1.3	1.3	5.9	8.8	30.9	60.3
German	dev	1889	1.2	1.2	1.7	4.6	1.1	42.5	56.3
	test	2765	1.2	1.4	1.4	4.0	1.8	44.1	54.1
	train	4633	1.2	1.3	1.5	4.4	10.3	38.9	50.7
Hindi	dev	1813	1.0	1.1	0.4	2.6	10.6	55.3	34.0
	test	2719	1.0	1.0	0.3	2.3	24.2	32.3	43.5
	train	4522	1.0	1.1	0.4	1.7	18.2	42.9	39.0
Indonesian	dev	1854	1.1	1.1	0.1	0.6	18.2	27.3	54.5
	test	2780	1.1	1.5	0.5	0.9	12.5	29.2	58.3
	train	4618	1.1	1.2	0.3	0.4	0.0	57.9	42.1
Italian	dev	1847	1.1	1.1	0.8	4.3	7.5	31.2	61.3
	test	2803	1.1	1.4	1.4	4.6	2.3	36.7	60.9
	train	4612	1.1	1.3	1.2	4.3	2.5	30.7	66.8
Japanese	dev	1781	0.5	0.9	3.6	1.5	14.8	51.9	33.3
	test	2652	0.5	0.9	2.3	1.6	14.0	41.9	44.2
	train	4410	0.5	1.4	5.3	1.8	10.0	32.5	57.5
Korean	dev	1844	0.6	1.3	7.9	4.2	6.4	42.3	51.3
	test	2750	0.5	0.9	5.4	3.6	13.0	51.0	36.0
	train	4601	0.5	1.0	6.0	2.6	13.4	46.2	40.3
Polish	dev	1829	1.0	1.1	0.8	4.4	7.4	38.3	54.3
	test	2766	1.1	1.1	0.6	3.3	7.7	38.5	53.8
	train	4589	1.1	1.3	1.5	3.7	4.8	35.1	60.1
Portuguese	dev	1837	1.1	1.4	2.3	3.5	6.2	41.5	52.3
	test	2769	1.1	1.1	0.8	3.6	5.0	30.7	64.4
	train	4586	1.1	1.1	0.8	3.0	9.4	38.1	52.5
Russian	dev	1855	1.1	1.4	2.4	6.6	13.0	30.1	56.9
	test	2769	1.1	1.2	2.0	6.1	5.3	39.6	55.0
	train	4588	1.1	1.5	2.9	5.6	6.6	41.7	51.7
Simplified Chinese	dev	1781	0.4	0.4	2.1	3.2	12.3	59.6	28.1
	test	2664	0.4	0.5	4.4	4.0	8.4	43.0	48.6
	train	4486	0.4	0.8	7.2	4.3	6.2	44.6	49.2
Spanish	dev	1836	1.2	2.3	7.6	4.6	4.7	24.7	70.6
	test	2756	1.3	2.7	10.5	3.4	7.4	22.1	70.5
	train	4602	1.3	2.5	10.3	3.3	7.2	34.0	58.8
Thai	dev	1807	0.9	1.1	0.7	2.5	13.3	53.3	33.3
	test	2699	1.0	1.9	3.0	3.8	22.5	26.5	51.0
	train	4550	0.9	1.1	0.7	2.8	39.7	24.6	35.7
Traditional Chinese	dev	1800	0.3	0.8	4.9	2.1	0.0	23.7	76.3
	test	2691	0.4	0.9	9.9	3.8	1.0	28.4	70.6
	train	4504	0.4	1.2	11.5	2.9	1.6	24.0	74.4
Turkish	dev	1832	1.0	1.5	1.9	2.9	11.1	38.9	50.0
	test	2767	1.0	1.1	0.7	3.0	8.3	35.7	56.0
	train	4592	1.0	1.1	1.4	4.0	12.1	34.6	53.3
Ukrainian	dev	1843	1.1	1.3	1.0	3.5	4.7	34.4	60.9
	test	2787	1.1	1.1	1.7	5.2	13.1	33.1	53.8
	train	4635	1.1	1.3	2.6	5.0	9.1	35.3	55.6
Vietnamese	dev	1873	1.1	1.2	1.1	4.3	3.7	30.9	65.4
	test	2815	1.1	1.3	1.2	3.2	7.8	33.3	58.9
	train	4671	1.1	1.3	1.1	3.1	7.7	33.6	58.7

Table 2: HAT dataset statistics (X→English).

Source Language	Split	NS	T2SLR	T2SLRHG	CHR (%)	SHR (%)	SHR (%) (0-5]	SHR (%) (5-25]	SHR (%) (25-100]
Arabic	dev	1930	1.4	1.4	3.9	20.4	23.4	56.7	19.8
	test	2901	1.4	1.5	3.7	19.9	22.5	57.3	20.2
	train	4822	1.4	1.4	3.0	17.1	19.2	63.2	17.7
Dutch	dev	1905	0.9	0.9	0.7	5.9	6.2	89.3	4.5
	test	2821	0.9	1.0	0.7	5.5	8.4	77.3	14.3
	train	4697	0.9	1.0	1.1	6.6	8.7	79.0	12.3
French	dev	1854	0.9	1.0	0.1	1.4	26.9	57.7	15.4
	test	2745	0.9	1.0	0.1	1.0	14.8	63.0	22.2
	train	4589	0.9	1.0	0.3	1.9	10.2	64.8	25.0
German	dev	1823	0.9	0.9	1.1	7.5	9.6	73.5	16.9
	test	2729	0.9	1.0	1.5	7.8	9.8	73.4	16.8
	train	4567	0.9	1.0	1.2	7.6	8.9	79.0	12.1
Hindi	dev	1983	1.1	1.2	1.5	4.6	28.6	49.5	22.0
	test	2969	1.1	1.1	1.0	5.4	21.2	52.5	26.2
	train	4950	1.1	1.2	1.2	4.3	26.4	55.2	18.4
Indonesian	dev	1646	0.9	1.0	0.1	1.6	11.5	88.5	0.0
	test	2462	0.9	1.0	0.1	1.0	32.0	64.0	4.0
	train	4130	0.9	1.0	0.0	0.5	9.5	90.5	0.0
Italian	dev	1919	0.9	0.9	0.8	6.9	12.0	68.4	19.5
	test	2868	1.0	1.0	1.3	6.2	11.3	69.5	19.2
	train	4783	0.9	1.0	0.9	7.1	12.0	70.4	17.6
Japanese	dev	1789	2.6	4.7	12.6	6.8	5.7	32.0	62.3
	test	2647	2.5	4.2	9.6	6.6	8.0	31.4	60.6
	train	4460	2.5	3.9	8.4	6.2	7.6	30.4	62.0
Korean	dev	1960	2.3	2.9	5.4	8.9	4.6	45.4	50.0
	test	2945	2.3	2.8	5.6	10.2	11.0	52.7	36.3
	train	4895	2.4	3.3	7.2	8.4	7.1	55.3	37.7
Polish	dev	1934	1.0	1.1	1.2	5.3	5.9	74.5	19.6
	test	2915	1.0	1.1	1.2	6.3	7.0	78.9	14.1
	train	4849	1.0	1.2	1.6	4.4	9.4	69.8	20.8
Portuguese	dev	1752	1.0	1.0	0.2	2.1	13.9	75.0	11.1
	test	2708	1.0	0.9	0.3	2.7	4.1	74.0	21.9
	train	4509	1.0	1.1	0.4	2.0	7.6	75.0	17.4
Russian	dev	1735	1.1	1.2	1.7	7.1	17.1	69.9	13.0
	test	2567	1.1	1.2	1.2	5.6	23.1	63.6	13.3
	train	4294	1.1	1.1	1.3	5.9	20.2	57.7	22.1
Simplified Chinese	dev	1832	4.1	4.1	5.7	10.0	39.1	31.0	29.9
	test	2717	4.1	4.6	8.9	11.3	41.2	21.8	37.0
	train	4545	4.1	4.4	8.3	9.8	32.4	30.1	37.5
Spanish	dev	1757	1.0	1.0	0.5	3.6	6.2	60.9	32.8
	test	2652	1.0	1.3	2.1	4.9	5.4	73.6	20.9
	train	4382	1.0	0.9	0.5	4.5	9.6	73.2	17.2
Thai	dev	1909	1.2	1.5	4.6	10.8	1.9	40.8	57.3
	test	2859	1.2	1.3	2.6	7.7	2.7	33.0	64.3
	train	4774	1.2	1.3	2.5	6.8	2.8	32.5	64.7
Traditional Chinese	dev	1892	3.5	3.9	5.0	7.1	22.2	35.6	42.2
	test	2875	3.6	3.8	5.3	8.0	27.5	34.5	38.0
	train	4763	3.7	4.2	6.6	7.7	18.7	42.8	38.5
Turkish	dev	1941	1.0	1.1	1.2	9.0	21.3	67.2	11.5
	test	2895	1.0	1.1	1.5	8.5	19.9	71.5	8.5
	train	4848	1.0	1.1	1.2	7.6	19.2	69.2	11.6
Ukrainian	dev	1961	1.1	1.1	0.9	7.2	14.2	76.6	9.2
	test	2943	1.1	1.1	0.8	5.8	21.1	71.3	7.6
	train	4882	1.1	1.1	0.8	5.5	12.3	77.7	10.0
Vietnamese	dev	1892	1.0	1.1	0.8	4.3	19.8	40.7	39.5
	test	2816	1.0	1.1	0.8	2.9	16.0	45.7	38.3
	train	4709	1.0	1.1	0.6	2.8	15.2	49.2	35.6

Table 3: HAT benchmark on F1 metric.

(a) English→X

Target Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.19	0.15	0.17	0.18	0.19	0.21	0.30	0.25	0.38	**0.41**
Dutch	0.16	0.07	0.13	0.18	0.15	0.23	0.26	0.26	**0.43**	0.40
French	0.11	0.19	0.12	0.36	0.34	0.39	0.45	0.47	0.52	**0.55**
German	0.15	0.15	0.19	0.29	0.23	0.30	0.35	0.37	**0.45**	**0.45**
Hindi	0.07	0.11	0.10	0.13	0.17	0.17	0.21	0.21	0.24	**0.28**
Indonesian	0.09	0.05	0.04	0.07	0.06	0.09	0.12	0.09	0.23	**0.24**
Italian	0.21	0.17	0.19	0.33	0.31	0.31	0.42	0.43	0.55	**0.58**
Japanese	0.11	0.17	0.13	0.13	0.16	0.17	0.18	0.22	**0.27**	0.23
Korean	0.18	0.22	0.23	0.21	0.21	0.28	0.29	0.30	**0.37**	0.33
Polish	0.12	0.15	0.16	0.21	0.20	0.23	0.29	0.32	**0.40**	**0.40**
Portuguese	0.21	0.18	0.23	0.29	0.27	0.31	0.38	0.43	**0.48**	0.47
Russian	0.23	0.27	0.27	0.36	0.34	0.36	0.43	0.47	0.51	**0.52**
Simplified Chinese	0.18	0.21	0.16	0.30	0.31	0.29	0.37	0.43	0.46	**0.52**
Spanish	0.29	0.21	0.21	0.31	0.29	0.32	0.39	0.43	0.53	**0.55**
Thai	0.21	0.20	0.09	0.19	0.22	0.21	0.29	0.28	0.35	**0.39**
Traditional Chinese	0.38	0.40	0.33	0.25	0.26	0.26	0.30	0.34	0.38	**0.43**
Turkish	0.05	0.09	0.13	0.18	0.17	0.23	0.26	0.24	**0.36**	0.33
Ukrainian	0.17	0.21	0.19	0.28	0.28	0.30	0.35	0.39	0.49	**0.52**
Vietnamese	0.08	0.14	0.13	0.18	0.16	0.18	0.21	0.23	0.27	**0.30**
Average	0.17	0.18	0.17	0.23	0.23	0.25	0.31	0.32	0.40	**0.42**

(b) X→English

Source Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.46	0.45	0.47	0.49	0.48	0.49	0.51	0.54	0.57	**0.61**
Dutch	0.18	0.18	0.24	0.29	0.27	0.31	0.39	0.44	0.54	**0.56**
French	0.03	0.00	0.05	0.06	0.07	0.08	0.09	0.11	0.11	**0.20**
German	0.21	0.23	0.26	0.37	0.37	0.38	0.50	0.52	**0.61**	0.53
Hindi	0.14	0.10	0.15	0.18	0.16	0.16	0.19	0.18	0.19	**0.20**
Indonesian	0.03	0.08	0.11	0.06	0.05	0.07	0.08	0.08	0.15	**0.17**
Italian	0.23	0.17	0.20	0.34	0.33	0.36	0.47	0.49	0.56	**0.62**
Japanese	0.33	0.27	0.27	0.28	0.27	0.34	0.38	0.37	0.43	**0.49**
Korean	0.24	0.32	0.32	0.32	0.33	0.34	0.36	0.40	0.42	**0.47**
Polish	0.18	0.22	0.26	0.29	0.29	0.34	0.39	0.39	0.49	**0.54**
Portuguese	0.15	0.10	0.14	0.18	0.18	0.18	0.25	0.30	0.33	**0.37**
Russian	0.19	0.20	0.22	0.27	0.25	0.30	0.37	0.35	0.37	**0.39**
Simplified Chinese	0.27	0.33	0.28	0.29	0.30	0.35	0.35	0.33	0.41	**0.42**
Spanish	0.12	0.09	0.19	0.27	0.28	0.31	0.36	0.40	0.44	**0.45**
Thai	0.20	0.24	0.20	0.27	0.24	0.27	0.30	0.28	**0.33**	**0.33**
Traditional Chinese	0.25	0.32	0.20	0.23	0.24	0.29	0.30	0.30	0.36	**0.37**
Turkish	0.24	0.25	0.28	0.33	0.30	0.32	0.40	0.41	**0.49**	0.46
Ukrainian	0.19	0.20	0.25	0.31	0.30	0.37	0.42	0.45	0.56	**0.58**
Vietnamese	0.09	0.13	0.10	0.12	0.10	0.13	0.14	0.15	0.19	**0.22**
Average	0.20	0.20	0.22	0.26	0.25	0.28	0.33	0.34	0.40	**0.42**

Table 4: HAT benchmark on Precision metric.

(a) English→X

Target Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.19	0.10	0.14	0.10	0.11	0.12	0.19	0.15	0.25	**0.28**
Dutch	0.14	0.08	0.09	0.10	0.09	0.13	0.16	0.16	**0.28**	**0.28**
French	0.24	0.13	0.27	0.23	0.22	0.25	0.32	0.34	0.40	**0.48**
German	0.12	0.12	0.14	0.18	0.15	0.19	0.25	0.27	0.34	**0.42**
Hindi	0.04	0.06	0.08	0.08	0.10	0.10	0.14	0.14	0.16	**0.21**
Indonesian	0.10	0.08	0.02	0.04	0.03	0.05	0.07	0.05	0.14	**0.16**
Italian	0.17	0.10	0.14	0.21	0.19	0.21	0.31	0.30	0.45	**0.52**
Japanese	**0.30**	0.14	0.10	0.07	0.09	0.10	0.11	0.14	0.17	0.15
Korean	0.19	0.24	**0.30**	0.13	0.13	0.17	0.19	0.20	0.27	0.26
Polish	0.08	0.09	0.10	0.12	0.11	0.14	0.18	0.20	0.27	**0.29**
Portuguese	0.15	0.12	0.17	0.18	0.16	0.19	0.26	0.30	**0.38**	**0.38**
Russian	0.15	0.18	0.19	0.23	0.22	0.23	0.31	0.34	0.39	**0.43**
Simplified Chinese	0.15	0.24	0.20	0.18	0.19	0.18	0.24	0.29	0.33	**0.40**
Spanish	0.25	0.14	0.21	0.19	0.18	0.20	0.26	0.29	0.38	**0.43**
Thai	0.16	0.15	0.06	0.12	0.13	0.13	0.20	0.19	0.27	**0.35**
Traditional Chinese	**0.56**	0.36	0.44	0.15	0.15	0.15	0.18	0.21	0.24	0.29
Turkish	0.04	0.05	0.09	0.10	0.10	0.14	0.16	0.15	**0.25**	**0.25**
Ukrainian	0.12	0.14	0.17	0.18	0.18	0.19	0.25	0.28	0.38	**0.43**
Vietnamese	0.05	0.08	0.08	0.11	0.09	0.11	0.13	0.14	0.19	**0.22**
Average	0.17	0.14	0.16	0.14	0.14	0.16	0.21	0.22	0.29	**0.33**

(b) X→English

Source Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.32	0.32	0.32	0.34	0.34	0.33	0.36	0.40	0.42	**0.48**
Dutch	0.11	0.10	0.14	0.18	0.17	0.20	0.26	0.30	0.39	**0.48**
French	0.02	0.00	0.02	0.03	0.03	0.04	0.05	0.06	0.06	**0.12**
German	0.13	0.14	0.17	0.25	0.25	0.28	0.41	0.42	0.55	**0.59**
Hindi	0.08	0.07	0.08	0.11	0.10	0.09	0.12	0.11	0.12	**0.13**
Indonesian	0.03	0.05	0.07	0.03	0.02	0.04	0.04	0.04	0.08	**0.10**
Italian	0.15	0.10	0.12	0.21	0.20	0.23	0.32	0.35	0.41	**0.52**
Japanese	**0.60**	0.24	0.21	0.17	0.17	0.21	0.25	0.24	0.30	0.36
Korean	**0.39**	0.25	0.37	0.20	0.21	0.22	0.24	0.27	0.29	0.34
Polish	0.15	0.14	0.17	0.18	0.18	0.22	0.26	0.27	0.34	**0.43**
Portuguese	0.11	0.06	0.08	0.10	0.10	0.10	0.15	0.19	0.22	**0.28**
Russian	0.12	0.12	0.14	0.16	0.15	0.19	0.24	0.23	0.25	**0.29**
Simplified Chinese	0.21	0.26	0.21	0.19	0.20	0.23	0.24	0.23	0.30	**0.32**
Spanish	0.12	0.08	0.12	0.16	0.17	0.19	0.24	0.28	0.32	**0.37**
Thai	0.11	0.15	0.12	0.18	0.15	0.18	0.21	0.19	0.23	**0.24**
Traditional Chinese	**0.37**	0.30	0.27	0.14	0.14	0.17	0.18	0.19	0.23	0.25
Turkish	0.14	0.17	0.17	0.21	0.19	0.20	0.27	0.30	0.36	**0.38**
Ukrainian	0.12	0.12	0.15	0.20	0.19	0.25	0.30	0.34	0.44	**0.56**
Vietnamese	0.06	0.08	0.05	0.06	0.05	0.07	0.08	0.08	0.11	**0.13**
Average	0.18	0.14	0.16	0.16	0.16	0.18	0.22	0.24	0.29	**0.34**

Table 5: HAT benchmark on Recall metric.

(a) English→X

Target Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.19	0.30	0.21	0.74	0.76	0.75	0.71	0.70	**0.82**	0.73
Dutch	0.18	0.05	0.27	0.80	0.75	0.82	0.62	0.75	**0.85**	0.69
French	0.07	0.32	0.08	0.79	0.75	**0.82**	0.74	0.75	0.74	0.66
German	0.19	0.23	0.30	**0.73**	0.54	0.72	0.57	0.59	0.63	0.49
Hindi	0.15	0.53	0.13	**0.56**	**0.56**	0.55	0.48	0.50	0.44	0.42
Indonesian	0.08	0.04	0.21	**0.71**	0.58	**0.71**	0.54	0.46	0.67	0.54
Italian	0.30	0.61	0.29	**0.80**	0.77	0.64	0.67	0.72	0.72	0.65
Japanese	0.07	0.23	0.16	0.60	0.60	**0.70**	0.56	0.58	0.56	0.47
Korean	0.18	0.21	0.18	0.57	0.49	**0.69**	0.59	0.54	0.58	0.45
Polish	0.23	0.51	0.46	**0.80**	0.69	**0.80**	0.67	0.74	0.76	0.65
Portuguese	0.35	0.35	0.35	**0.75**	0.71	0.74	0.69	0.73	0.66	0.61
Russian	0.49	0.53	0.47	**0.83**	0.75	**0.83**	0.71	0.76	0.76	0.67
Simplified Chinese	0.25	0.19	0.13	**0.83**	0.82	0.81	0.80	0.81	0.79	0.75
Spanish	0.33	0.40	0.21	**0.88**	**0.88**	0.87	0.81	0.85	0.87	0.74
Thai	0.29	0.27	0.21	**0.56**	0.54	0.53	0.48	0.51	0.50	0.46
Traditional Chinese	0.29	0.45	0.26	**0.92**	0.88	0.87	0.81	0.88	0.90	0.84
Turkish	0.07	0.60	0.20	0.65	0.64	**0.76**	0.65	0.60	0.63	0.50
Ukrainian	0.31	0.48	0.22	0.68	0.63	**0.73**	0.59	0.64	0.72	0.64
Vietnamese	0.29	0.60	0.39	**0.72**	0.61	0.64	0.52	0.61	0.47	0.49
Average	0.23	0.36	0.25	0.73	0.68	**0.74**	0.64	0.67	0.69	0.60

(b) X→English

Source Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.79	0.77	0.85	0.89	0.80	**0.90**	0.88	0.84	**0.90**	0.83
Dutch	0.44	0.83	0.73	0.80	0.79	0.70	0.76	0.80	**0.86**	0.68
French	0.11	0.00	0.70	**0.74**	0.70	**0.74**	0.67	0.63	0.52	0.59
German	0.56	**0.77**	0.63	0.71	0.74	0.58	0.63	0.69	0.67	0.49
Hindi	0.65	0.24	**0.86**	0.55	0.48	0.52	0.52	0.45	0.49	0.42
Indonesian	0.04	0.32	0.20	0.60	0.48	**0.84**	0.76	0.44	**0.84**	0.60
Italian	0.45	0.44	0.55	**0.86**	0.83	0.81	0.85	0.83	0.85	0.76
Japanese	0.23	0.32	0.37	0.75	0.70	**0.88**	0.77	0.76	0.77	0.73
Korean	0.17	0.45	0.28	0.78	0.78	**0.81**	0.75	0.80	0.80	0.74
Polish	0.23	0.51	0.51	0.79	0.78	0.81	0.78	0.73	**0.86**	0.73
Portuguese	0.23	0.27	0.44	**0.77**	0.74	0.67	0.66	0.73	0.71	0.53
Russian	0.52	0.59	0.45	0.79	0.76	0.69	**0.80**	0.71	0.72	0.58
Simplified Chinese	0.36	0.47	0.42	0.63	0.63	**0.76**	0.68	0.57	0.67	0.60
Spanish	0.12	0.10	0.40	**0.83**	0.75	0.74	0.73	0.74	0.73	0.59
Thai	**0.68**	0.58	0.62	0.62	0.56	0.58	0.51	0.53	0.54	0.50
Traditional Chinese	0.18	0.34	0.17	0.74	0.73	**0.88**	0.80	0.71	0.79	0.72
Turkish	0.72	0.52	0.75	**0.80**	0.72	0.77	0.78	0.67	0.78	0.59
Ukrainian	0.47	0.51	0.71	0.70	0.73	0.68	0.71	0.68	**0.76**	0.60
Vietnamese	0.15	0.33	0.53	0.67	0.63	**0.74**	0.65	0.59	0.67	0.60
Average	0.37	0.44	0.54	**0.74**	0.70	**0.74**	0.72	0.68	0.73	0.63

Table 6: HAT benchmark on MCC metric.

(a) English→X

Target Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.17	0.14	0.13	0.21	0.22	0.24	0.32	0.28	**0.42**	**0.42**
Dutch	0.10	0.11	0.12	0.26	0.22	0.30	0.29	0.32	**0.48**	0.42
French	0.11	−0.01	0.23	0.38	0.35	0.40	0.45	0.47	0.51	**0.53**
German	0.10	0.19	0.15	0.31	0.23	0.32	0.34	0.36	**0.44**	0.43
Hindi	0.07	0.14	0.08	0.16	0.20	0.20	0.22	0.23	0.24	**0.28**
Indonesian	0.08	0.08	0.04	0.13	0.11	0.17	0.18	0.13	**0.29**	0.28
Italian	0.21	0.18	0.19	0.36	0.33	0.31	0.42	0.43	0.54	**0.56**
Japanese	0.14	0.25	0.11	0.18	0.21	0.23	0.22	0.26	**0.29**	0.25
Korean	0.33	0.19	0.21	0.22	0.21	0.30	0.30	0.29	**0.37**	0.31
Polish	0.09	0.17	0.15	0.27	0.23	0.28	0.31	0.35	**0.43**	0.41
Portuguese	0.19	0.20	0.20	0.32	0.29	0.34	0.39	0.44	**0.48**	0.46
Russian	0.19	0.25	0.23	0.37	0.34	0.38	0.42	0.46	**0.50**	**0.50**
Simplified Chinese	0.26	0.16	0.14	0.34	0.35	0.33	0.40	0.45	0.48	**0.52**
Spanish	0.27	0.20	0.27	0.37	0.36	0.38	0.43	0.47	**0.56**	0.55
Thai	0.17	0.12	0.04	0.19	0.21	0.20	0.27	0.27	0.33	**0.37**
Traditional Chinese	0.39	0.38	0.33	0.32	0.31	0.32	0.34	0.39	0.43	**0.46**
Turkish	0.01	0.19	0.12	0.21	0.20	0.28	0.29	0.26	**0.37**	0.32
Ukrainian	0.09	0.16	0.15	0.29	0.27	0.31	0.33	0.37	0.48	**0.49**
Vietnamese	0.04	0.13	0.13	0.22	0.18	0.21	0.22	0.25	0.26	**0.29**
Average	0.16	0.17	0.16	0.27	0.25	0.29	0.32	0.34	**0.42**	0.41

(b) X→English

Source Language	BLASER-2.0-QE	XCOMET-XXL-QE	MetricX24-Hybrid-XXL-QE-BF16	Qwen3-8B	GPT-OSS-20B	DS-V3.1	Qwen3-235B-A22B	GPT-OSS-120B	Gemini-2.5-Flash	Gemini-2.5-Pro
Arabic	0.31	0.29	0.33	0.37	0.34	0.36	0.39	0.43	0.48	**0.51**
Dutch	0.15	0.18	0.23	0.31	0.29	0.31	0.40	0.45	**0.54**	**0.54**
French	0.08	0.08	0.07	0.13	0.13	0.14	0.16	0.17	0.16	**0.25**
German	0.13	0.19	0.23	0.34	0.35	0.33	0.46	0.48	**0.57**	0.50
Hindi	0.09	−0.01	0.13	0.14	0.12	0.11	0.15	0.13	**0.16**	0.15
Indonesian	0.09	0.09	0.10	0.10	0.07	0.15	0.16	0.11	**0.24**	0.23
Italian	0.18	0.13	0.20	0.36	0.34	0.37	0.48	0.50	0.56	**0.60**
Japanese	0.34	0.25	0.19	0.27	0.25	0.36	0.38	0.37	0.43	**0.47**
Korean	0.24	0.23	0.26	0.26	0.28	0.29	0.31	0.36	0.39	**0.42**
Polish	0.20	0.17	0.25	0.30	0.29	0.35	0.39	0.38	0.49	**0.52**
Portuguese	0.12	0.12	0.16	0.24	0.23	0.22	0.28	0.34	**0.37**	0.36
Russian	0.15	0.17	0.17	0.28	0.26	0.29	**0.38**	0.35	0.37	0.36
Simplified Chinese	0.20	0.28	0.16	0.19	0.20	0.28	0.27	0.24	**0.34**	**0.34**
Spanish	0.12	0.16	0.17	0.31	0.30	0.32	0.37	0.41	**0.44**	0.43
Thai	0.15	0.18	0.13	0.22	0.18	0.22	0.24	0.22	**0.27**	**0.27**
Traditional Chinese	0.27	0.22	0.16	0.18	0.19	0.28	0.28	0.27	**0.35**	0.34
Turkish	0.17	0.21	0.24	0.31	0.26	0.29	0.38	0.37	**0.46**	0.41
Ukrainian	0.17	0.17	0.25	0.30	0.30	0.36	0.41	0.44	**0.55**	**0.55**
Vietnamese	0.06	0.11	0.09	0.14	0.10	0.17	0.17	0.17	0.23	**0.24**
Average	0.17	0.17	0.19	0.25	0.24	0.27	0.32	0.33	**0.39**	**0.39**

Table 7: Annotation guideline examples illustrating Mistranslation vs. Hallucination

Source	Reference Translation	Given Translation	Error Category	Explanation
Este es el árbol más alto del bosque	This is the tallest tree in the forest	This is a large tree in the forest	Mistranslation	"large", "big" can be associated with "tall" in terms of size
Este es el árbol más alto del bosque	This is the tallest tree in the forest	This is the tallest and oldest tree in the forest	**Hallucination**	"and oldest" is an extra information not present in the source
Siga recto y luego gire a la derecha	Go straight and then take a right turn	Go straight and then take a correct turn	Mistranslation	"right" and "correct" can have same meaning/semantic in isolation
Siga recto y luego gire a la derecha	Go straight and then take a right turn	From here, first go straight and then take a right turn	**Hallucination**	"From here, first" is an extra information not present in the source
Me debes $10	You owe me $10	You owe me $100	Mistranslation	"100" represents number which can be associated with number "10" in source text
Me debes $10	You owe me $10	You owe me $10 million	**Hallucination**	"million" is an extra information
él es un hombre valiente	He is a brave man	He is not a brave man	**Hallucination**	"not" is an extra information not present in source
Le atropelló un coche.	He was hit by a car.	He was hit by a car a car.	**Hallucination**	"a car" is repeated in the translation but not in the source

Figure 8: LLM prompt used for hallucination detection task.

You are an expert in translation quality assessment. Please assess the following translated text to determine if it contains hallucinations. A hallucination consists in a translation error that is completely unrelated to the text in the source language, adding or altering the meaning of the translation greatly, in a way that the error does not bear any resemblance with the source. A mistranslation on the other hand, is a translation error that does not always alter the meaning of the source, and when it does, it still can have some resemblance to the meaning or can be associated with the source text. Your job is to only identify hallucination error and ignore mistranslation error. If you are in doubt if a span is a Hallucination or Mistranslation, please don't consider it as hallucination. You should also consider unnecessary incorrect repetitions in the translation as hallucinations. However, do not consider as hallucinations repeated words or phrases that are a correct translation of source. Focus on additions or fabrications, but ignore omissions, minor rephrasing, cultural adaptations, or fluency improvements that preserve the original meaning.
You can ask yourself the following three questions to determine if the error is a hallucination:
1. Does any source text span belong to the common meaning category of the erroneous translation span?
2. Does any source text span have a semantic connection with this translation span irrespective of the context?
3. Can you try to come up with a reasonable theory on how any source text span can be associated with this translation span?
If the answer is yes to any of the above question , then the error is a Mistranslation, and if not, then it is Hallucination.

Please take a look at the below examples to differentiate mistranslation vs hallucination. The examples here are for illustrative purpose in a few sample languages, actual task may contain a different language.When a tranlsation contains hallucination then "Label" field in the below examples is set to 1 else 0.
Example 1:
Source: He was hit by a car.
Translation: Le atropelló un coche un coche.
Label: 1
Reason: "un coche" (a car) is repeated in the translation but not in the source.

Example 2:
Source: I want to say it to his face.
Translation: Quiero decírselo cara a cara.
Label: 0
Reason: "cara a cara" is a valid expresion in Spanish.

Example 3:
Source: Este es el árbol más alto del bosque
Translation: This is the tallest and oldest tree in the forest
Label: 1
Reason: "and oldest" is an extra information not present in the source

Example 4:
Source: Este es el árbol más alto del bosque
Translation: This is a large tree in the forest

Label: 0
Reason: "large", "big" can be associated with "alto" in terms of size

Example 5:
Source: La aplicación falló debido a un error.
Translation: The application crashed due to a insect
Label: 0
Reason: "insect" and "bug" can have same meaning/semantic in isolation

Example 6:
Source: La aplicación falló debido a un error.
Translation: The application in my phone crashed due to a bug
Label: 1
Reason: "in my phone" is an extra information not present in the source

Example 7:
Source: You owe me 10 dollar
Translation: Me debes 100 dollar
Label: 0
Reason: "10 dollar" incorrectly translation to "100 dollar", not a hallucination

Example 8:
Source: You owe me 10 dollar
Translation: Me debes 100 millones de dólares
Label: 1
Reason: "millones" is extra information added in translation

Example 9:
Source: El no es un hombre valiente
Translation: He is a brave man
Label: 0
Reason: "not" is missing in translation and it is an omission error not hallucination

Example 10:
Source: El es un hombre valiente
Translation: He is not a brave man
Label: 1
Reason: "not" is an extra information not present in source

Provide only the assessment in the specified format, and do not include any additional comments, formatting, or chattiness.

source locale: {source_locale}
target locale: {target_locale}
source text: {source_text}
target text: {target_text}

Please respond with this exact format:
{
Reason: Provide a brief reasoning explaining your assessment.

Label: Output true if hallucinations are present, or false if none are detected.
Score: Output a float number between 0 and 1 that represents the probability of hallucination or the confidence level.
}

Figure 9: Baseline LLM prompt used in ablation study.

You are an expert in translation quality assessment. Please assess the following translated text to determine if it contains hallucinations, where a hallucination is defined as the addition of information, facts, or details in the target text that are not present or implied in the source text, or a significant distortion of the meaning beyond faithful translation. Focus on additions or fabrications, but ignore omissions, minor rephrasings, cultural adaptations, or fluency improvements that preserve the original meaning. Please note: Base your assessment solely on semantic comparison without external knowledge or assumptions.
Provide only the assessment in the specified format, and do not include any additional comments, formatting, or chattiness.

source locale: {source_locale}
target locale: {target_locale}
source text: {source_text}
target text: {target_text}

Please respond with this exact format:
{
Reason: Provide a brief reasoning explaining your assessment.
Label: Output true if hallucinations are present, or false if none are detected.
Score: Output a float number between 0 and 1 that represents the probability of hallucination or the confidence level.
}