When Translations Surprise: Human Awareness of Predictability in Translations

Cristian García-Romero[†] **Miquel Esplà-Gomis**[†‡] **Felipe Sánchez-Martínez**[†‡]
[†]Dep. de Llenguatges i Sistemes Informàtics
[‡]Institut Universitari d'Investigació Informàtica
Universitat d'Alacant (Spain)
{cristian.gr,miquel.espla,fsanchez}@ua.es

Abstract

Machine translation (MT) has achieved near-human quality for some language pairs, yet its output remains distinct from human translation, primarily in its predictability. While MT systems generate low-perplexity text, humans produce less predictable outputs. This raises the question of whether humans can intuitively use this difference in predictability to distinguish between human- and machine-translated text. We report on a study with 30 native Spanish speakers tasked with identifying the origin of English-to-Spanish translations. We compared their performance against two perplexity-based baselines: a large language model capturing fluency, and a neural MT model, conditioned on the source text, capturing both fluency and adequacy. Our findings reveal that human judgments correlate with fluency-based perplexity, but show no correlation with the perplexity that also accounts for adequacy. This suggests that annotators' decisions are driven by the target text's fluency. Consequently, a simple computational baseline using source-aware perplexity significantly outperforms human annotators. This work contributes to a deeper understanding of human perception of MT, highlighting a potential bias in current evaluation protocols toward fluency over adequacy. This bias may lead to an overestimation of the capabilities of highly fluent systems and underscores the need for evaluation methods ensuring translation adequacy is not overlooked.

Keywords: machine translation detection, machine-generated text detection, human vs. machine-generated translations, human evaluation, translation perplexity

1. Introduction

Translation is a core activity in cross-linguistic communication, carried out both by human translators and by machine translation (MT) systems. Today, state-of-the-art MT systems, typically built on the Transformer neural architecture (Vaswani et al., 2017), achieve performance levels that approach human quality for some language pairs in shared tasks such as the 2024 Conference on Machine Translation (WMT) (Kocmi et al., 2024).[1]

While the quality of MT output is often judged by human preferences (Clark et al., 2021), this evaluation does not fully capture whether machine-generated text aligns with the type of translations produced by humans. When we look at translations as final products, human and machine outputs remain distinct (Vanmassenhove et al., 2019; Roberts et al., 2020; Luo et al., 2024). Because MT systems are trained on large-scale bilingual data, they tend to generate highly predictable translations, often characterized by low perplexity. Human translators, by contrast, go beyond fluency and adequacy: they attend to nuance, register, stylistic choices, and pragmatic appropriateness, resulting in outputs that are less predictable. This is illustrated in Figure 1, which shows the per-word perplexity of machine-generated and human translations computed using the NLLB-200 3.3B multilingual translation model (NLLB Team et al., 2024). As can be observed, human translations are generally less predictable, resulting in higher per-word perplexity compared to machine-generated translations.

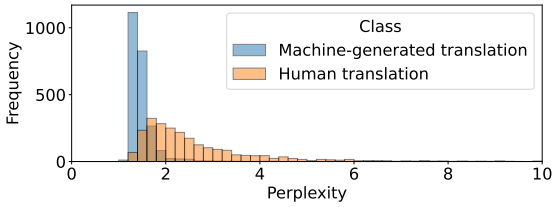


Figure 1: Histograms of the per-word perplexity obtained with the NLLB-200 3.3B NMT model (PPL_{NMT}) for the human and machine-generated translations in our dataset (see Section 2.2). MT-generated translations were obtained with MADLAD (Kudugunta et al., 2023) and similar trends are observed with the rest of machine-generated translations.

Previous work has shown that humans struggle to differentiate between human translations (HT) and state-of-the-art machine-generated translations (Calvo-Ferrer, 2024), a difficulty also observed when assessing text generated with other advanced AI systems (Dugan et al., 2023). However, these studies do not account for the pre-

[1] https://www2.statmt.org/wmt24/translation-task.html

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8615–8627
11–16 May 2026.

dictability of the translations under evaluation. Consequently, it remains unclear whether translators rely on their intuition about linguistic predictability —if such intuition exists— to distinguish human translations from those produced by state-of-the-art neural MT systems.

In this paper, we report the results of a study with 30 native Spanish speakers, both with and without formal translation training, which were asked to identify the origin of English–Spanish translations. Our evaluation included translations produced by multilingual and bilingual neural MT systems, models implementing the full Transformer architecture (encoder and decoder), and systems based on large language models (LLMs).

To determine whether predictability is somehow perceived (and leveraged) by human evaluators, we also computed the perplexity of translations using two different models: a large-language model (LLM) and a multilingual neural MT (NMT) model. The key distinction between these two models is that the LLM is monolingual; it only considers the target-language text and thus primarily assesses translation fluency. In contrast, the NMT model is bilingual; it also takes the source text into account and therefore evaluates both fluency and adequacy relative to the source.

Our findings reveal a partial correlation between the perplexity computed using an LLM on the monolingual target-language text and human judgments. Paradoxically, the perplexity computed using the bilingual NMT model shows no such correlation with human judgments, despite being a much better predictor for automatic detection of machine-translated text. This suggests that human annotators may be focusing too heavily on the fluency of the target-language text (which the monolingual LLM assesses), while largely disregarding adequacy with respect to the source-language text (which the bilingual NMT model assesses). This focus on fluency could be negatively impacting their overall performance.

We hope this finding contributes to a better understanding of how translators perceive and evaluate MT outputs and informs the design of evaluation protocols that move beyond surface-level judgments, and avoid the potential bias towards fluency over adequacy that current human evaluation protocols may have. More broadly, it highlights the need to examine which linguistic signals, if any, remain reliable indicators of human translation.

The rest of the paper is organized as follows. Section 2 describes the experimental setup, whereas Section 3 presents and discusses the results. Then we present a description of related work in Section 4. The paper ends with some concluding remarks and ethical considerations, followed by a discussion of the limitations of our study.

2. Experimental Setting

This section details the human evaluation carried out to distinguish HT from MT-generated translations (Section 2.1), describes the datasets used (Section 2.2), and introduces the use of perplexity as a proxy for predictability (Section 2.3).

2.1. Human Evaluation

Annotators. The evaluation involved 30 native Spanish-speaking participants: 15 with training in translation (final-year students in Translation Studies; annotators 1-14 and 16), and 15 without such training (the remaining participants). All participants were also proficient in English to evaluate the translated content effectively.

These participants were organized into 10 groups of three annotators each to ensure that a small, identical set of sentences was assigned for annotation. Each annotator was assigned an individual data package (see Section 2.2). Groups sharing sentences are numbered consecutively (e.g., 1–3 or 4–6). Groups 4 and 5 were exceptions to the main grouping: Group 4 included one annotator without translation training, and Group 5 included two.

For each instance in our data set, we presented annotators with a source segment in English, and two translations in Spanish, one HT and one MT. Annotators were then asked to identify the HT. This task aims at measuring how good humans are in identifying MT-generated content when it is compared to genuine human translations. Furthermore, this setup facilitates a direct comparison of human annotators' performance against an automatic approach based on the perplexity obtained from different text-generation models.

Evaluation Platform. We conducted the human evaluation using KEOPS (Ramírez-Sánchez et al., 2020),[2] a web-based tool for manual evaluation of parallel corpora. We modified KEOPS to fit our specific purposes. We customized the ranking mode of the tool and instructed the annotators to assign a rank of "1" to the HT and a rank of "2" to the MT.[3] We also updated the descriptions within the tool to align with the objectives of the task. The modified KEOPS interface, as used for the task, is displayed in Figure 2.

2.2. Datasets

We build on the test splits released by the WMT news translation shared tasks for the language pair

[2] https://github.com/paracrawl/keops

[3] Annotations that assigned the same rank to both translations were subsequently removed from our data.

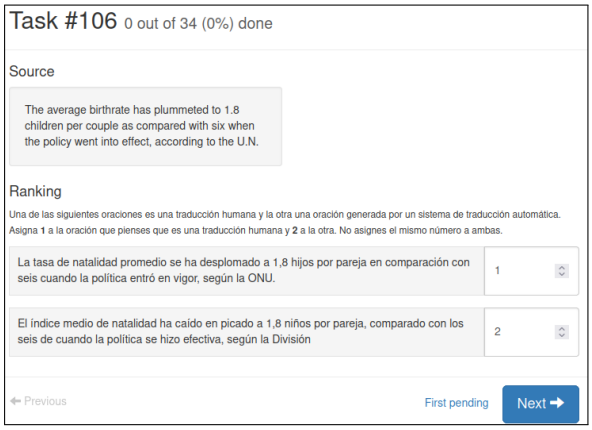


Figure 2: KEOPS interface used by our annotators.

MT system	Evaluation metric		
	COMET	BLEU	chrF2
Opus	85.2	26.8	55.3
MADLAD	86.5	28.5	56.7
Tower	86.8	28.4	56.2

Table 1: Evaluation on the FLORES+ devtest split of the MT systems used.

English–Spanish (2008–2013).[4] This dataset consists of sentence pairs where the source segments are original, and the target segments are HT.

We extend this dataset by adding translations from state-of-the-art MT systems: MADLAD-400 (Kudugunta et al., 2023) (hereafter MADLAD or MADL), Opus-MT (Tiedemann and Thottingal, 2020) (hereafter Opus) and Tower Instruct (Alves et al., 2024) (hereafter Tower).[5] MADLAD and Opus implement an encoder-decoder Transformer model (Vaswani et al., 2017); MADLAD is multilingual whereas Opus is bilingual. Tower is a multilingual, instruction-tuned, decoder-only model.

Table 1 reports the automatic MT evaluation quality metrics BLEU (Papineni et al., 2002), COMET (Rei et al., 2022),[6] and chrF2 (Popović, 2015) for the MT systems used. These figures were calculated on the FLORES+ (NLLB Team et al., 2024) devtest split for the English–Spanish language pair. The results indicate that the differences between MADLAD and Tower are quite small, with Opus slightly underperforming both. The three systems report state-of-the-art results as of October 2025.

[4] https://www.statmt.org/wmt13/translation-task.html

[5] We use the prompt recommended by the authors: https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2

[6] https://huggingface.co/Unbabel/wmt22-comet-da

WMT edition	Number of source sentences		
	Before filtering	After filtering	After filtering (%)
2008	349	292	83.66
2009	370	314	84.86
2010	505	435	86.13
2011	598	480	80.26
2012	604	480	79.47
2013	500	388	77.60
Total	2926	2389	81.65

Table 2: Number of source sentences per WMT split. Each source sentence is paired with its HT and the MT-generated translations.

The dataset was constructed by pairing each source sentence with its HT and the MT-generated translations from all MT system; subsequently, any entry containing a duplicated target sentence was discarded. Table 2 reports the number of source sentences per WMT edition before and after the filtering process.

We created individual data packages from the resulting dataset for each annotator. Each package contained 12 randomly selected unique entries and 5 additional entries shared across packages for groups of three annotators to measure inter-annotator agreement. We selected one MT system's translation per entry. While these MT-generated translations were generally chosen at random, we ensured that the same specific MT outputs were used for the 5 overlapping entries across the annotators within each group.

2.3. Measurement of Predictability

We use per-word perplexity (PPL) computed from the logits produced by: (a) a surrogate multilingual translation model, NLLB-200 3.3B (NLLB Team et al., 2024); and (b) a surrogate multilingual pretrained (non-instruction-tuned) LLM, LLaMA 3.1 8B (Grattafiori et al., 2024). These metrics are used as a means to measure the predictability of both human and machine-generated translations. Equation 1 shows how the PPL based on a surrogate translation model (PPL_{NMT}) is computed for the target segment t with T tokens, conditioned on the source sentence s, by feeding t to the model through teacher-forcing, rather than by decoding from the probability distribution of the model:

$$\text{PPL}_{\text{NMT}} = \exp\Big(-\frac{1}{T}\sum_{i=1}^{T}\log \text{P}(t_i \mid t_{<i}, s)\Big), \quad (1)$$
$$\text{P}(t_i \mid t_{<i}, s) = \text{softmax}(\text{logits}_i)[t_i],$$

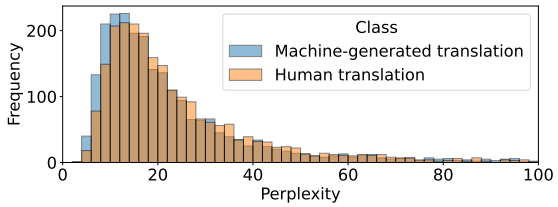


Figure 3: Histograms of the per-word perplexity obtained with the LLaMA 3.1 LLM model (PPL_{LLM}) for the human and machine-generated translations in our dataset (see Section 2.2). MT-generated translations were obtained with MADLAD (Kudugunta et al., 2023) and similar trends are observed with the rest of machine-generated translations.

The computation of the PPL based on an LLM (PPL_{LLM}) is done analogously after removing the conditioning on the source segment s.

As Figure 1 shows, MT outputs consistently exhibit lower PPL_{NMT} than HT. While the figure specifically plots the PPL_{NMT} for MADLAD, similar trends hold for the remaining machine-generated translations. In contrast, from the perspective of an LLM, HT and MT outputs are quite similar in terms of perplexity. This is evident in Figure 3, which shows the PPL_{LLM} for translations produced by MADLAD (a pattern that generalizes to the other MT systems).

3. Results and Discussion

In this section, we report the annotators' accuracy in determining which is the HT when the source text and two possible translations —HT and MT— are provided. Inspired by Mitchell et al. (2023), we compare human results to those obtained by two simple baselines that assume that the HT is usually less predictable and, therefore, exhibits higher perplexity (analogous to Mitchell et al. (2023), who use entropy as a baseline to identify AI-generated text): one that computes the perplexity according to a surrogate neural MT model (PPL_{NMT}) and one that does it by using an LLM (PPL_{LLM}). We also explore the relation between PPL_{NMT} and PPL_{LLM} and the performance of annotators, and try to figure out whether annotators may have any nuance regarding the concept of predictability of a translation and the chances that it has been produced by a human or an MT system.

In our experiments with human annotators, each annotator classified 12 unique entries, plus 5 additional entries shared with two other annotators in their group. Due to limitations of the evaluation platform (see Section 2.1), some annotations were invalid and had to be removed from the results. Specifically, one annotation sample was removed for annotators 19, 20, 21, and 30, as well as for the overlapping group 19–21.

Data split		Opus	MADL	Tower
All	Human	†62.8	†52.7	†45.5
	PPL_{NMT}	95.1	98.2	71.4
	PPL_{LLM}	46.5	65.1	63.0
Experts	Human	63.7	53.9	†53.3
	PPL_{NMT}	96.7	97.8	74.7
	PPL_{LLM}	55.0	69.7	73.3
Non-experts	Human	†62.0	51.3	38.0
	PPL_{NMT}	93.5	98.8	68.4
	PPL_{LLM}	38.0	60.0	53.2

Table 3: Accuracy obtained by human annotators and by the two perplexity-based baselines for the two data splits (*experts* and *non-experts*) and for the combination of all the data (*all*). Scores whose difference from the score immediately above is statistically significant are underlined. † indicates statistically significant differences between Human and PPL_{LLM}.

3.1. Performance of Human Annotators vs. PPL-based Baselines

Table 3 reports the annotators' accuracy in determining which of the two translations is the one produced by a human, across three different data splits and the three MT systems. The table also reports the results obtained by the two baselines described above, as well as the outcomes of statistical significance tests computed via approximate randomization (Noreen, 1989) (10,000 iterations; $p < 0.05$) to identify statistically significant pairwise differences across all systems. Since different data was provided to different annotator groups, the data splits are categorized by annotator background: *Experts* includes samples from annotators with translation training; *Non-experts* includes samples from those with no translation training; and *All* represents the combined data from the other two splits.

Overall, human classification accuracy across the different MT systems is close to the chance level (i.e., 50%), although annotators with translation training achieve slightly higher accuracies. Among the three MT systems used, Opus translations appear marginally easier for annotators to correctly identify, whereas MADLAD translations are closest to random classification for both experts and non-experts (i.e., small deviation from chance). In contrast, Tower translations are significantly more difficult to detect, particularly for non-expert annotators. These trends suggest that the human annotators find the translations generated by Opus —a bilingual NMT system— less human-like than those produced by MADLAD —a multilin-

Annotators	Pairwise (%)	Krippendorff's α
1–3	73.3	0.44
4–6	33.3	-0.30
7–9	60.0	0.16
10–12	60.0	0.25
13–15	46.7	-0.04
16–18	86.7	0.72
19–21	50.0	0.08
22–24	33.3	-0.30
25–27	60.0	0.25
28–30	46.7	0.00
All	55.1	0.08

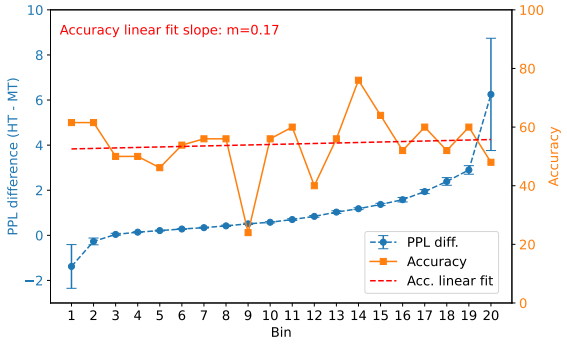
Table 4: Inter-annotator agreement on the data packages shared across different groups of annotators. Annotators 1-14 and 16 have training in translation while the remaining annotators do not.

gual NMT system—, and that both are perceived less human-like than the translations generated by Tower —a multilingual LLM.

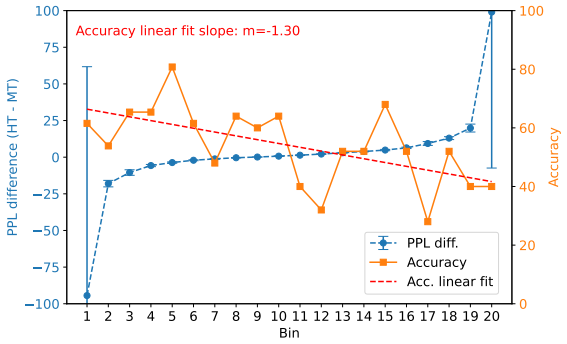
Remarkably, despite its simplicity, the PPL_{NMT} baseline outperforms both groups of human annotators by a large statistically-significant margin in all cases. It achieves near-perfect classification results for Opus and MADLAD. While its performance on Tower is lower, the PPL_{NMT} baseline still remains substantially higher than the accuracy achieved by the human annotators.

As regards the PPL_{LLM} baseline, its performance is much closer to human annotators. Namely, this baseline obtains results that outperform human annotators for MADLAD and Tower MT systems, although statistical significance tests results are inconsistent across groups of human annotators. Surprisingly, the results obtained for Opus are substantially lower, and statistically-significant worse than human annotators (the expert annotators are the exception to the statistically significant differences). Nevertheless, this baseline is still far from the results obtained with the PPL_{NMT} baseline, which consistently outperforms PPL_{LLM} across all evaluated cases, with statistically significant differences for Opus and MADLAD.

Inter-Annotator Agreement. Table 4 reports per-group results in terms of inter-annotator agreement, using both pairwise agreement and Krippendorff's alpha. We observe that the two metrics are strongly positively correlated across the individual samples (excluding aggregated results), as measured by Pearson's linear correlation: $r = 0.9125, p < 0.0005$. Overall, annotator agreement is generally weak, with most values being close to chance levels (i.e., 50% for pairwise inter-annotator agreement and 0 for Krippendorff's alpha). Interestingly, groups with translation training (annotators 1-14 and 16) achieve higher levels of agreement, whereas groups without translation training (remaining annotators) exhibit more variability and noise, with no clear trend.



(a) Using PPL_{NMT}.



(b) Using PPL_{LLM}.

Figure 4: Annotators' accuracy (orange lines) and differences in PPL between HT and MT (blue lines).

3.2. Do Human Annotators Perceive (and Leverage) Translation Predictability?

To investigate how human annotators may be using the predictability of the translations as a cue for detecting the HT, we processed the data points used to compute the accuracy scores in Table 3: we sorted them in ascending order of the difference in PPL between the HT and the corresponding MT shown to the annotators. We then grouped them into 20 bins, with each bin containing 25 data points but covering a different range of PPL differences. This was done both for PPL_{NMT} and PPL_{LLM}.

Figure 4 displays the average and standard deviation of the PPL differences in each bin (shown in blue), alongside the annotators' accuracy in identifying HTs (shown in orange) both for PPL_{NMT} and for PPL_{LLM}. As can be seen in Figure 4a, the magnitude of the PPL_{NMT} difference between the human translations and their MT-produced counterparts does not significantly impact the annotators' accuracy. This is confirmed by the Kendall correlation

	PPL_{NMT}(HT)-PPL_{NMT}(MT)		PPL_{LLM}(HT)-PPL_{LLM}(MT)	
Data split	Kendall's τ	p-value	Kendall's τ	p-value
All	0.04	0.79	-0.42	0.01
Experts	-0.21	0.22	-0.43	0.01
Non-experts	0.25	0.13	-0.29	0.08

Table 5: Kendall correlation between the difference of the per-word perplexity (PPL) computed for the HT and the MT, and the accuracy of the annotators.

computed between the PPL_{NMT} obtained on the HT and on the MT and shown in Table 5; no clear correlation is observed neither for annotators with training in translation and for non-experts. A manual inspection of 50 samples from the experts split and another 50 from the non-experts split, corresponding to the highest difference in PPL_{NMT} (i.e., pairs of sentences in the bins on the right of Figure 4a), showed no severe hallucinations between the source sentences and their machine-generated translations that could have affected the results reported above.

The results obtained on the PPL_{LLM} lead to different conclusions. In this case, Figure 4b reflects that accuracy decreases as the difference between the PPL_{LLM} computed on the HT and the MT grows. In other words, when the PPL_{LLM} of the HT is clearly higher than the one of the MT, it seems to become more difficult for annotators to identify which is the human translation, while when it is the other way round, it seems to be much easier. The correlation between the difference of the PPL_{LLM} computed on HT and MT-generated translations and human accuracy is also reported in Table 5. As can be seen, this correlation is identified for both experts and non-experts, although it is slightly weaker for the last ones.

These results seem to indicate that the process followed by human annotators is somehow more related to the predictability of the text in the target language, without getting much influence by the source text. We can safely conclude that annotators do not seem to assume that human translations are less predictable than MT-generated translations; they seem to annotate the translation with the lower PPL_{LLM} as the human translation.

Surprisingly, according to the results reported in Table 3, the predictability of the translation, taking into account both the source and the target text, as computed with the PPL_{NMT}, seems to be the best option to differentiate HT and MT-generated translations. These results open up the question of whether humans, with the appropriate training, would be capable of perceiving this type of predictability and, if so, which is the best way to instruct them to leverage it for the task of differentiating HT from MT-generated translations.

4. Related Work

Several studies have explored differences between human and machine-generated translations (Vanmassenhove et al., 2019; Roberts et al., 2020; Luo et al., 2024). However, modern MT systems now produce highly convincing texts, making the HT vs. MT classification task increasingly challenging —evidence suggests that higher-quality MT outputs are harder to detect automatically (Aharoni et al., 2014). As a result, research on MT detection has evolved along two main lines: automatic detection methods, which aim to train models able to differentiate HT from MT, and human annotation, which assess how well humans can perform the same task. In what follows, we review prior work in both areas.

4.1. Automatic Detection

Early work on MT detection focused on identifying outputs from statistical MT systems, leveraging fluency and linguistic features (Arase and Zhou, 2013; Li et al., 2015). With the emergence of NMT, Nguyen-Son and Echizen (2018) were among the first to address the detection of neural outputs, focusing on distinguishing MT from original (non-translated) text using n-gram-based fluency and noise features. Much of the subsequent research has focused on sentence-level classification, primarily motivated by the fact that NMT systems are typically trained at this level, despite emerging trends toward coarser granularity (Kocmi et al., 2024).

Approaches vary from feature-based models and recurrent neural networks to pre-trained monolingual and multilingual transformers. For instance, Bhardwaj et al. (2020) compared multiple modeling paradigms across several English–French domains, while Fu and Nederhof (2021) studied lexical diversity via n-gram statistics and BERT-based classifiers for English translations from multiple source languages. Alternative formulations have also been explored: Nguyen-Son et al. (2021) proposed iteratively back-translating texts to measure stability differences between MT and human-generated text.

As sentence-level classification becomes in-

creasingly difficult for high-quality MT —and particularly for short segments (Bhardwaj et al., 2020; Nguyen-Son et al., 2021)— several studies have turned to paragraph- and document-level detection. Nguyen-Son et al. (2017) exploited Zipfian distributions at the document level, while Nguyen-Son et al. (2018, 2019) investigated coherence-based features. van der Werff et al. (2022) compared document- and sentence-level classification and found that detection of the former substantially outperforms the latter for German–English, using both support vector machines and pre-trained transformer classifiers. Extending this line of work, Chichirau et al. (2023) evaluated multilingual transformer-based models across seven source languages for English translations. More recently, Chen et al. (2025) proposed combining a surrogate speech model with a monolingual encoder-based language model to separate original, human-produced text (non-translated) from MT-generated text.

Overall, automatic detection research has evolved from feature-based methods toward neural approaches that exploit larger context and multilingual representations. Yet, even state-of-the-art detectors often degrade in performance as MT quality improves, motivating complementary studies of human ability to perform the same task — particularly given that human judgment is typically considered the gold standard for MT evaluation.

4.2. Human Annotation

Compared to automatic HT vs. MT classification, human detection has received substantially less attention, and we find limited research works on this topic. Popel et al. (2020) describe the model that won the WMT18 (Bojar et al., 2018) translation shared task on the news domain for Czech–English. They also conducted a human evaluation to distinguish between HT and MT outputs from their system and Google Translate. Expert translators, MT researchers, and other evaluators were asked to annotate the texts. Results showed that their MT system was not significantly distinguished from HT, whereas Google Translate outputs were more easily identified. In both cases, annotators exhibited high variance and difficulty in the task. More recently, Calvo-Ferrer (2024) evaluated human ability to distinguish human- and machine-generated subtitles for English–Spanish, using ChatGPT (an instruction-tuned model) rather than traditional NMT. A coarser granularity was explored, as entire episodes were translated at once rather than individual sentences. They observed similar difficulties, with annotators struggling to identify MT.

Other works in adjacent fields, such as AI-generated text detection, further support these findings. For instance, Gehrmann et al. (2019) introduced GLTR, a tool that helps non-expert annotators identify machine-generated text by showing statistics, such as token-level probabilities, leveraging a surrogate language model. Ippolito et al. (2020) explored the role of decoding algorithms, finding that this decision affects detection performance. They also report that non-expert annotators easily detect semantic errors, whereas automatic detectors focus on statistical anomalies. Clark et al. (2021) investigated training methods for non-expert annotators but found limited improvements. Dugan et al. (2023) analyzed detection across domains, model sizes, and decoding strategies, focusing on boundary-detection rather than binary classification. Across all these studies, humans exhibit difficulties in detecting machine-generated text. Interestingly, performance can improve when annotators are provided with statistics from the model (Gehrmann et al., 2019), incentives (Clark et al., 2021), or detailed instructions on identifying machine-generated content (Dugan et al., 2023).

Building on these prior studies, our work fills gaps in the literature by comparing neural and LLM state-of-the-art MT systems. We provide insights into differences in human classification of HT vs. MT across these paradigms, and also analyze how the predictability of the evaluated translations correlates with annotators' classification performance. We focus on the news domain, a widely studied and challenging benchmark for human detection due to the high quality and conventional style of translations. The quality of MT systems in this domain has steadily improved, driven by continuous evaluation campaigns such as those organized as part of the WMT conferences on MT. This setup allows us to assess human ability to distinguish HT from MT in a realistic, high-quality context.

5. Concluding Remarks

Predictability is a key feature for detecting machine-generated text, yet prior works have not explored if human annotators are able to leverage their intuition about predictability to distinguish between HT and MT. In this study, we investigate this question by using per-word perplexity as a proxy for predictability, computed from both an LLM (PPL_{LLM}) and an NMT model (PPL_{NMT}). The LLM processes only the translation, capturing fluency, whereas the NMT model is also fed with the source text, capturing both fluency and adequacy. Our results show that annotators' performance correlates with PPL_{LLM}, but no correlation is found with PPL_{NMT}. This suggests that human judgments are primarily driven by the fluency of the target translations, with no clear evidence that the source sentence systematically

influences their decisions.

Surprisingly, a simple baseline that assigns HT and MT labels based on the highest and lowest PPL_{NMT} values, respectively, outperforms human annotators by a wide margin. In contrast, using PPL_{LLM} as the predictor yields substantially worse performance, highlighting that the source sentence contains critical information for detecting machine-generated translations.

Consistent with prior work, we find that distinguishing HT from state-of-the-art MT is a challenging task for human annotators, particularly in the English–Spanish news domain. The combination of strong performance from the PPL_{NMT} baseline, the correlation between human judgments and PPL_{LLM}, and the absence of correlation with PPL_{NMT} strongly suggests that annotators may overlook the source sentence when evaluating translations. As a consequence, current human MT evaluation protocols may be systematically biased toward fluency over adequacy. A relevant example is the widely adopted direct assessment (DA) protocol —partly used in the most recent WMT general translation shared task (Kocmi et al., 2025)— which, if such a bias exists, could lead to an overestimation of the translation capabilities of highly fluent LLMs compared to traditional encoder-decoder NMT systems. This concern is particularly relevant given that LLMs can be highly sensitive to prompts, and some studies have shown that certain prompts can cause the model to produce largely inadequate translations (Zhang et al., 2023; Zhu et al., 2024), meaning that the source text is effectively ignored—even though the output may still appear fluent. However, unlike our setup, DA typically presents annotators with the source sentence and a single MT output for scoring—and, in some cases, the corresponding HT as a reference. These differences in task configuration and experimental design mean that our findings, while indicative of a potential bias toward fluency over adequacy, should be interpreted with caution.

Future work should explore evaluation settings with richer context, more diverse language pairs, and improved annotator training, as well as hybrid protocols that combine human and automatic signals. While some of these approaches have been explored in AI-generated content detection (Gehrmann et al., 2019; Clark et al., 2021), our findings suggest that human behavior in MT evaluation —particularly the reliance on fluency over source adequacy— may differ from general AI-generated text detection, and therefore conclusions from prior work may not fully generalize to the MT domain.

6. Ethical Considerations

All annotators were fully informed that they were participating in an academic study aimed at analyzing human performance in distinguishing HT from MT. Participation was voluntary, and no monetary compensation was provided. Particular care was taken to ensure participant anonymity and prevent any collection or disclosure of personally identifiable information.

For transparency and reproducibility, we release both the code and data used in this study to facilitate further analysis and validation by the research community.[7]

7. Limitations

While our study provides insights into human performance in distinguishing HT and MT, several limitations should be considered. We focus on a single language pair and domain —English–Spanish news— which may limit the generalizability of our findings. Different language pairs or domains could reveal distinct patterns in human annotation behavior. Additionally, our analysis is conducted at the sentence level, whereas recent trends in machine translation evaluation, such as recent WMT editions, are shifting toward coarser granularities like paragraphs or entire documents. Evaluating larger contexts could uncover trends not captured at the sentence level.

The experimental design also results in each annotator working on a different subset of sentences. While this prevents strict one-to-one comparisons, it allows us to examine broader trends in classification when results are aggregated, which aligns with the focus of this work. Furthermore, although each annotator received a relatively small number of sentences, this setup ensures that the task could be completed in a reasonable amount of time, reducing the risk of careless or random responses caused by fatigue or decreased interest in the task.

For the annotator assignment process, we divided the participants into two groups: experts and non-experts. The non-expert group is heterogeneous, while the expert group is relatively narrow, consisting of final-year translation students. Other studies have included professional translators, and incorporating such participants —or exploring alternative group configurations— could provide additional insights and potentially reveal different patterns of performance.

Finally, we acknowledge some limitations in the surrogate models used in our study. To compute PPL as a proxy for fluency, we used LLaMA 3.1,

[7] https://github.com/transducens/human-eval-wmt-en-es.

and to assess both fluency and adequacy, we leveraged NLLB. First, we did not evaluate alternative models or different model sizes. More importantly, although NLLB reports having filtered WMT data from its training dataset (NLLB Team et al., 2024), the authors of LLaMA 3.1 have not disclosed the specific sources of its training data, only mentioning that much of their data was from the web (Grattafiori et al., 2024). This raises the possibility of data contamination. If such contamination were present, we might expect the PPL_{LLM} pattern observed in Figure 3 to be narrower for human translations (i.e., the original texts potentially included in web data), similar to the pattern seen for machine translations in Figure 1 for PPL_{NMT}, since the model's pre-training objective encourages lower perplexity on training data. However, in the absence of detailed information about the data used for training, we cannot confirm this possibility.

These limitations highlight opportunities for future work, including exploring multiple language pairs and domains, evaluating larger granularity levels, and considering alternative configurations of human annotator groups, in order to improve our understanding of human judgment in distinguishing HT and MT. Additionally, future work could also investigate other models as surrogates, ideally allowing assessment of potential data contamination.

Acknowledgements

We thank Alicia Martínez Ochoa for her assistance in coordinating the human annotators' participation in this study.

Work funded by the Spanish Ministry of Science, Innovation and Universities, the Spanish Research Agency (MICIU/AEI/10.13039/501100011033/) and the European Regional Development Fund A way to make Europe (ERDF) through R+D+i project PID2024-158157OB-C31, and by Universitat d'Alacant through project GRE23-08A. Cristian García-Romero is funded by Generalitat Valenciana and the European Social Fund through the research grant CIACIF/2021/365. Some of the computational resources used were funded by the Valencian Government and the European Regional Development Fund (ERDF) through project IDIFEDER/2020/003.

8. Bibliographical References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–295, Baltimore, Maryland. Association for Computational Linguistics.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1597–1607, Sofia, Bulgaria. Association for Computational Linguistics.

Shivendra Bhardwaj, David Alfonso Hermelo, Phillippe Langlais, Gabriel Bernier-Colborne, Cyril Goutte, and Michel Simard. 2020. Human or neural translation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6553–6564, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

José Ramón Calvo-Ferrer. 2024. Can you tell the difference? a study of human vs machine-translated subtitles. *Perspectives: Studies in Translation Theory and Practice*, 32(6):1115–1132.

Yongjian Chen, Mireia Farrús, and Antonio Toral. 2025. The potential of speech features to discriminate between original and machine-translated texts. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Malina Chichirau, Rik van Noord, and Antonio Toral. 2023. Automatic discrimination of human and neural machine translation in multilingual scenarios. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 217–226, Tampere, Finland. European Association for Machine Translation.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12763–12771.

Yingxue Fu and Mark-Jan Nederhof. 2021. Automatic classification of human translation and machine translation: A study from the perspective of lexical diversity. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 91–99, online. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. 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, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 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.

Yitong Li, Rui Wang, and Hai Zhao. 2015. A machine learning method to distinguish machine translation from human translation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 354–360, Shanghai, China.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Hoang-Quoc Nguyen-Son and Isao Echizen. 2018. Detecting computer-generated text using fluency and noise features. In *International Conference of the Pacific Association for Computational Linguistics*, pages 288–300, Singapore. Springer Singapore.

Hoang-Quoc Nguyen-Son, Huy H. Nguyen, Ngoc-Dung T. Tieu, Junichi Yamagishi, and Isao Echizen. 2018. Identifying computer-translated paragraphs using coherence features. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Hoang-Quoc Nguyen-Son, Tran Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. 2021. Machine translated text detection through text similarity with round-trip translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5792–5797, Online. Association for Computational Linguistics.

Hoang-Quoc Nguyen-Son, Tran Phuong Thao, Seira Hidano, and Shinsaku Kiyomoto. 2019. Detecting machine-translated paragraphs by matching similar words. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 521–532. Springer.

Hoang-Quoc Nguyen-Son, Ngoc-Dung T. Tieu, Huy H. Nguyen, Junichi Yamagishi, and Isao Echi

Zen. 2017. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1504–1511. IEEE.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, New York.

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.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 32(1).

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.

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

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C. Lipton. 2020. Decoding and diversity in machine translation. In *Proceedings of the Resistance AI Workshop at 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, CA.

Tobias van der Werff, Rik van Noord, and Antonio Toral. 2022. Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, USA.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

9. Language Resource References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin,

Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva,

Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.