Natural Language Processing (2026), **32**, pp. 589–637
doi:10.1017/nlp.2026.10041

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Lexical diversity and translation literalness in machine translation: a comparative study of pretrained, fine-tuned, and few-shot approaches

Zeynep Yirmibeşoğlu Balal and Tunga Güngör

Computer Engineering, Boğaziçi University, Türkiye
Corresponding author: Zeynep Yirmibeşoğlu Balal; Email: zeynep.yirmibesoglu@std.bogazici.edu.tr

(Received 14 October 2025 UTC; revised 27 May 2026 UTC; accepted 16 August 2026 UTC;
first published online 30 September 2026 UTC)

Abstract

This paper presents a comparative study of neural machine translation (MT) models and large language models, emphasizing translation literalness, lexical diversity, and translation quality. We evaluate open-source neural machine translation (NMT) models, LLaMa 3.1 Instruct, and GPT-4o across nine language pairs in pretrained, fine-tuned, and few-shot prompting settings. This comprehensive analysis underscores the differential strengths of each model, contributing valuable insights into their respective utility in MT. Our findings in the literary and news domains show that fine-tuning reduces translation literalness, allowing freer translations across models, most strongly for LLaMa, which produces the least literal translations across domains. The effect of fine-tuning on lexical diversity is model-dependent: it decreases consistently for LLaMa but remains stable or is improved for NMT. Few-shot prompting, on the other hand, enhances GPT-4o's lexical diversity across domains and significantly boosts its translation quality in the literary domain, yielding the highest overall COMET scores. Furthermore, our correlation analysis reveals a distinct domain effect: literary translation evaluation rewards diversity, whereas news translation evaluation favors literalness. Overall, the results highlight complementary strengths: few-shot GPT-4o for high-quality, lexically rich translations with low error rates, and fine-tuned LLaMa for freer, less literal translations.

Keywords: Machine translation; natural language generation; evaluation

1. Introduction

Over recent years, the field of machine translation (MT) has witnessed significant advancements, primarily driven by the introduction of the transformer architecture. This architecture enabled encoder–decoder neural machine translation (NMT) models to achieve notable success in translating a wide array of languages by capturing complex linguistic patterns and longer context. The next big break occurred with the emergence of large language models (LLMs), which started to transform the landscape of natural language processing (NLP), demonstrating superior capabilities in various tasks including translation. As LLMs gain increasing prominence for their ability to understand and generate text, it is essential to conduct a comparative analysis to understand their strengths and weaknesses relative to traditional NMT models. This paper endeavors to explore these differences by systematically assessing the performance of both types of systems, thus shedding light on their respective utilities in the ever-evolving domain of MT.

The evaluation of MT performance is predominantly conducted through quantitative metrics such as BLEU, COMET, and BLEURT scores, which provide standardized measures of translation

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

accuracy and fluency. Beyond these established criteria, we delve into two more dimensions: lexical richness and translation literalness. By examining these additional aspects, we aim to capture a more comprehensive picture of a MT model's ability to be creative in lexical choices and to generate freer translations.

We specifically focus on three models which are (i) the open-source OPUS-MT transformer model as representative of the NMT model and two LLMs as (ii) the open-source LLaMa 3.1 Instruct model with 8 billion parameters and (iii) the commercial GPT-4o model. By fine-tuning both NMT and LLM types of models, we aim to delve into the effects of fine-tuning on translation literalness, lexical diversity, and overall quality. In addition to this, we examine the efficacy of zero-shot and few-shot prompting for the LLMs to determine how well these models can handle translation tasks without extensive training. Our analysis spans nine language pairs as Spanish-English, Italian-English, German-English, French-English, Turkish-English, Russian-English, Chinese-English, Japanese-English, and Hindi-English, where we evaluate translation performance in both directions for each, and two domains as literary and news. Through these comprehensive experiments, we seek to understand each model's strengths and weaknesses across various translation scenarios. Specifically, this study addresses the following research questions:

- **RQ1 (Adaptation effects):** How do fine-tuning and in-context learning affect the lexical diversity and translation literalness aspects of MT models compared to their pretrained or zero-shot counterparts?
- **RQ2 (Quality–diversity trade-offs):** What are the trade-offs between translation quality, lexical diversity, and translation literalness across different model families (NMT vs. LLMs) and prompting strategies?
- **RQ3 (Domain differences):** How do these dynamics in quality, diversity, and literalness vary between different text domains, specifically literary and news texts?

Our contributions are as follows:

- To the best of our knowledge, this is the first study to compare lexical diversity or translation literalness in zero-shot, few-shot, and fine-tuned settings of both NMT models and LLMs.
- We adopt a **three-dimensional evaluation framework**—translation quality, lexical diversity, and translation literalness—that enables a more comprehensive comparison of NMT models and LLMs.
- We position **in-context learning**, particularly few-shot prompting, as an effective and practical alternative to fine-tuning for leveraging LLMs in translation, highlighting its advantages for translation quality and lexical diversity.
- We contribute insights into the role of **fine-tuning in open-source models**, showing its potential for producing less literal, more creative translations, especially relevant for literary contexts.
- We provide a **systematic correlation analysis** of lexical diversity, translation literalness, and translation quality metrics, clarifying how surface-based and semantic-based metrics capture different aspects of translation quality across domains.
- We incorporate a detailed **error analysis**, revealing that in-context learning with GPT-4o achieves the best balance of high translation quality and low error rates across both domains.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

- We identify a **clear domain effect** in translation evaluation, establishing that literary translation rewards lexical diversity and creativity, while news translation rewards literalness and accuracy.
- We benchmark model performance against **human translations**, revealing that while humans excel in reference-based metrics, nonreference metrics like COMET-Kiwi sometimes favor MT outputs (particularly GPT-4o), highlighting a metric preference for fluency over stylistic variation.

2. Related work

2.1 Large language models in translation

Recent work has established the effectiveness of LLMs for MT across many language pairs, particularly for high-resource languages. The presence of LLMs in this field is becoming more and more dominant. The Eighth Conference on Machine Translation (WMT23) (Kocmi *et al.* 2023) received very few submissions using LLMs, whereas LLMs took part in a majority of the systems in the Ninth Conference on Machine Translation (WMT24) (Kocmi *et al.* 2024a) and LLM-based submissions dominated the top rankings, often outperforming classical NMT approaches. The findings of WMT24 suggested near-perfect translations in some language pairs such as English-Spanish. The best-performing system in general was found to be Claude 3.5-Sonnet.

However, limitations persist: Hendy *et al.* (2023) noted that models such as ChatGPT and GPT-3.5 perform well for high-resource language pairs, but pointed out notably degraded translation quality with low-resource languages, a deduction confirmed by Jiao *et al.* (2023), with the additional observation that GPT-4 exhibits fewer hallucinations and mistranslation errors compared to GPT-3.5.

Parallel lines of work have explored how best to leverage LLMs for translation. Prompting strategies—including zero-shot and few-shot in-context learning—allow LLMs to translate without explicit fine-tuning. For instance, Garcia *et al.* (2023) show that only a handful of carefully chosen examples at inference time can enable an LLM to match specialized supervised models and commercial translation systems. Approaches to example selection, such as retrieving contextually similar parallel sentences as in relevance-based sampling of translation examples (fuzzy matches), have also been proposed to stabilize few-shot performance (Xu *et al.* 2020). Moslem *et al.* (2023) explore the usage of fuzzy matches for domain adaptation and find that in-context learning can surpass encoder-–decoder NMT models, especially for high-resource languages. Tang *et al.* (2025) enrich in-context examples by introducing adaptive few-shot prompting (AFSP), which automatically retrieves semantically similar translation examples from a large bilingual corpus using LLM embeddings. By selecting top-k demonstrations that closely match each input sentence and generating multiple output candidates and reranking them with a smaller language model, AFSP produces more relevant in-context examples and improves translation consistency in low-resource settings. In a similar spirit, Frontull and Ströhle (2025) propose fragment-shot prompting. They segment each source sentence into syntactic fragments and retrieve aligned translation fragments from a small parallel corpus. This syntactic-covering retrieval significantly enhances translations between two dialects of Ladin, showing how targeted few-shot samples can compensate for data scarcity. In sum, recent prompting advances continue to push the boundaries of MT with LLMs in low-resource scenarios and for domain adaptation tasks.

Fine-tuning and parameter-efficient adaptation of LLMs also yield high translation quality. Xu *et al.* (2024) experimented with fine-tuning on their LLaMa-2-based ALMA model (Touvron *et al.* 2023) and achieved remarkable improvements of over 12 BLEU and 12 COMET points across multiple translation directions. Their two-stage fine-tuning approach, utilizing monolingual data followed by high-quality parallel data, proved effective even with relatively smaller

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

models (7B or 13B parameters). With LLMs getting larger, the GPU demands for fully fine-tuning a model has become more challenging to fulfill. Because of this, parameter efficient fine-tuning (PEFT) mechanisms such as LoRA (Hu *et al.* 2021) and QLoRA (Dettmers *et al.* 2023) were proposed, which allow very large models to be adapted by training far fewer parameters. Alves *et al.* (2023) demonstrate that adapter-based fine-tuning (LoRA) can match full fine-tuning performance while updating only a small fraction of parameters. Zhang *et al.* (2023) leveraged QLoRA fine-tuning for sentence and document-level translation, where they observed significant BLEU score improvement over prompting by fine-tuning only 0.77% of the model parameters. Notably, though, fine-tuning can weaken an LLM's inherent flexibility: Alves *et al.* (2023) and Stap *et al.* (2024) find that models fine-tuned for translation exhibit reduced few-shot adaptability. This "fine-tuning paradox" suggests a trade-off between optimizing in-domain accuracy and retaining LLM generality.

Su *et al.* (2024) systematically compare PEFT methods (e.g. LoRA, adapters, prefix tuning) on low-resource MT benchmarks. They report that 6 out of 8 PEFT variants outperform full fine-tuning in both out-domain and in-domain experiments, with the Houlsby+Inversion adapter variant yielding the best overall BLEU scores. In practice, these methods offer strong trade-offs: updating fewer parameters drastically reduces computation effort and often preserves the model's general knowledge (mitigating forgetting), but may slightly limit peak performance if not carefully tuned. Indeed, Song *et al.* (2025) find that even fine-tuning on tiny supervised datasets ($\approx$ 1% of typical training size) can significantly close the gap on low-resource languages, whereas naive LoRA alone often underperforms full fine-tuning. Thus, while PEFT enables scalable adaptation of LLMs for MT, achieving the best results may still require some full fine-tuning or distillation. Researchers continue to balance these trade-offs: PEFT and prompt-based approaches extend LLMs to new languages and new domains with modest cost, whereas full fine-tuning yields peak quality at higher expense.

2.2 Lexical diversity and translation literalness

Assessing lexical richness (lexical diversity) helps identify whether an MT system over-relies on repetitive and simplistic language, which can lead to monotonous and less informative translations. This is particularly significant in domains like literature, where diverse expression and stylistic variety are essential to preserve the artistic quality and intent of the original work.

Vanmassenhove *et al.* (2019) compared the lexical richness of the outputs of NMT systems to those of SMT systems and human translation. They used BLEU and translation error rate (TER) to assess translation quality, and Yule's I, type-token ratio (TTR), and measure of lexical diversity (MTLD) to evaluate lexical richness. Their findings showed that recurrent neural network (RNN) and transformer MT models cannot match the lexical diversity of human translations, and that RNNs perform the worst in this aspect when compared to SMT and transformer systems. Vanmassenhove *et al.* (2021) confirm this finding after comparing human-generated translations to SMT, LSTM, and transformer NMT models.

In parallel, Freitag *et al.* (2022) examined the naturalness of MT outputs. They argue that translations can be accurate and fluent yet still sound less natural than native target language text due to translationese effects and reduced lexical diversity. They introduce a tagging method that contrasts natural versus translated data, biasing NMT systems toward producing more natural style translations. Their results show improved lexical richness and human preference for the tagged system's outputs, even if this comes at a small cost to translation adequacy.

Martínez *et al.* (2024) assessed the lexical diversity of conversational LLMs in English and how model parameters affect the diversity metrics. They made use of the metrics root TTR (RTTR), Maas, moving average TTR (MATTR), and MTLD. They observed that ChatGPT-4 has higher lexical diversity than ChatGPT-3.5 in most evaluation cases. Regarding the language model parameters, they found that lexical richness also slightly increases when the temperature and

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

presence penalty are increased separately. Reviriego *et al.* (2024) confirm that ChatGPT-3.5's outputs still tend to use fewer distinct words and exhibit lower lexical richness than human-generated translations, but with ChatGPT-4 the lexical diversity is increased and in some cases is greater than human translations. Likewise, Ploeger *et al.* (2024) note that MTs are generally lexically poorer than human translations, especially in literary contexts, and they propose a reranking approach to recover the lost lexical variety by using a classifier to rank hypotheses according to how much they resemble an originally written text. These studies underline that MT outputs often lack the rich vocabulary of human translations, suggesting that special interventions (e.g., reranking or creative post-editing) may be needed to approach human-like lexical richness.

Translation literalness is another aspect that can be evaluated in MT outputs. Literalness can be defined as high similarity in word order or lexical choice between the source and target language texts. Dankers *et al.* (2022) observed compositional literal translation behavior in transformer NMT models, specifically for idioms. Raunak *et al.* (2023) proposed two metrics to measure the literalness of translations that take into account the alignment of words and the monotonicity in word order. Their findings suggested that GPT models (*davinci-002* and *davinci-003*) produce less literal translations when compared to Microsoft Translator and a transformer encoder–decoder NMT model.

Baziotis *et al.* (2023) proposed the literal TER (LitTER) metric that automatically measures the frequency of literal translation errors. Using this metric, Stap *et al.* (2024) showed that fine-tuned models generate noticeably more nonliteral (creative) translations than the base models. Together, these findings imply that without careful tuning or guidance NMT and LLM translation systems tend to default to more literal renditions, and that specialized evaluation and training techniques are needed to encourage more varied, nonliteral outputs. Studies of idiomatic expressions reinforce this trend. Yang *et al.* (2025) find that even GPT-4 (the best system) fails on about 28% of Chinese idioms, often rendering them literally or incorrectly, highlighting persistent challenges for non-literal translation. They instruct-tune Qwen2.5 models to improve detection of idiom translation errors.

Building on these insights, recent work has sought to directly reduce overly literal translations. Huang *et al.* (2024) introduce DUAT (Difficult words Understanding Aligned Translation), a method that explicitly elicits general contextual understanding of difficult source words before generating the translation. They show that DUAT significantly reduces translation literalness (by 25–51%) while improving overall translation quality.

3. Background and definitions

3.1 Lexical diversity

Lexical diversity is a crucial aspect of natural language that is used to evaluate the quality of MT outputs (Bentivogli *et al.* 2016; Vanmassenhove *et al.* 2019). It reflects the variety of words in a text and provides insights into how well a translation model captures the richness of vocabulary. In the context of MT, assessing lexical diversity helps determine whether a model tends to repeat certain words or utilizes a broader lexicon, which can influence both fluency and adequacy.

We explore three prominent metrics used to measure lexical diversity: Type-token ratio (TTR), Yule's I, and measure of textual lexical diversity (MTLD). These metrics, although similar in their goal of evaluating word variety, differ in their sensitivity to text length and distribution of word frequencies. Each metric is explained briefly below and the formulation is provided in Appendix A.

Type-token ratio (TTR) is calculated as the ratio of the number of unique words (types) to the total number of words (tokens) in a text (Templin, 1957). One limitation of TTR is its sensitivity to text length, making it less reliable while comparing texts of different lengths.

Yule's I is based on the frequency distribution of words within a text and is less sensitive to text length than the TTR metric (Yule, 1944).

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Measure of textual lexical diversity (MTLD) calculates the average length of sequences of words that maintain a certain threshold (also called a type-token ratio threshold (McCarthy, 2005)) of lexical diversity. Among these metrics, MTLD is considered the most robust to text length. We use the same TTR threshold (0.72) as Vanmassenhove *et al.* (2021).

3.2 Translation literalness

Translation literalness is an aspect that affects the quality and accuracy of translated texts. Literal translation, which involves a word-for-word approach, may fail to convey the true meaning of the source language, particularly when dealing with idiomatic expressions or culturally specific references. This aspect may especially be important in the context of literary translation. While assessing translation quality, we evaluate how well MT models preserve the original meaning while maintaining fluency and naturalness.

To measure translation literalness in MT outputs, we use two metrics proposed by Raunak *et al.* (2023) as explained below. We use *awesome-align*[a] to compute word alignments.

Unaligned source words (USW) is used to differentiate translations with similar fluency and adequacy by measuring their literalness. This is done by aligning the source words with the target words and counting how many source words remain unaligned. When translation quality is controlled, less literal translations tend to have more USW.

Translation nonmonotonicity (NM) is a measure of translation literalness based on how closely the word order in the translation matches the source. It uses the nonmonotonicity metric introduced in Schioppa *et al.* (2021), which calculates deviations from the diagonal in word-to-word alignment, reflecting how much the word order changes.

3.3 Translation quality

We use the BLEU (Papineni *et al.* 2002) and COMET (Rei *et al.* 2020) metrics as the standard evaluation metrics used in MT. In addition to these two metrics, we follow the suggestion of Kocmi *et al.* (2024b) to include COMET-Kiwi (Rei *et al.* 2022) within our evaluation metrics, since it is not affected by references. They also recommend using an automatic evaluation metric with a different model assumption than COMET, such as BLEURT (Sellam *et al.* 2020). Thus, we use the BLEU, BLEURT-20, COMET, and nonreference-based COMET-Kiwi[b] for analyzing translation quality in this work. We use *sacrebleu* (Post 2018) to measure the BLEU scores. In addition to quality evaluation, we also conduct error analysis on translations with XCOMET[c] (Guerreiro *et al.* 2024).

4. Methodology

In this study, we aim to evaluate and compare the translation capabilities of NMT models and LLMs by analyzing translation quality, lexical diversity, and translation literalness. To achieve this, we make use of three types of models:

(1) Open-source NMT model: OPUS-MT transformer model
(2) Open-source large language model: LLaMa 3.1 Instruct model with 8B parameters
(3) Commercial large language model: GPT-4o

[a] https://github.com/vyraun/awesome-align
[b] https://huggingface.co/Unbabel/wmt22-cometkiwi-da
[c] https://huggingface.co/Unbabel/XCOMET-XXL

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

We fine-tune each model on a parallel dataset and compare with the pretrained (base) model. Additionally, we explore zero-shot and few-shot prompting with the LLMs to inspect how well the models handle the translation task without fine-tuning. By employing these different approaches, this paper seeks to capture the relative strengths and weaknesses of each model in various scenarios, revealing how they perform in terms of fidelity to the source text, diversity of lexical choice, and overall translation performance.

We conduct all the experiments in the literary domain for seven language pairs as Spanish-English, Italian-English, German-English, French-English, Turkish-English, Russian-English, and Chinese-English in both translation directions, and in the news domain for nine language pairs which are the aforementioned seven language pairs plus Hindi-English and Japanese-English in both translation directions. The code and the dataset splits are available on Github[d] except for the Turkish-English dataset, which is not publicly available due to copyright issues.

4.1 Fine-tuning

We fine-tune the three types of models with different settings, where we intend to assess the learning capabilities of the models and how lexical richness and literalness of the text respond to fine-tuning.

4.1.1 Neural machine translation models

The OPUS-MT transformer models were bilingually pretrained as part of the OPUS-MT project[e] on the OPUS corpus (Tiedemann 2012) and (when available for the language pair) the Tatoeba Translation Challenge corpus (Tiedemann 2020). We fine-tune the models for all language pairs in both directions and in two domains separately for five epochs on the entire training set of the specific translation direction. We use the best model according to the validation loss among the five epochs. We use a batch size of *16*, learning rate of 2×10^{-5}, and a weight decay of *0.01* for the literary domain. We do additional hyperparameter tuning for the news domain, since for some language pairs fine-tuning did not improve translation quality. We believe this is due to abundance of news data in the pre-training data of the models. The specific models used for each translation direction and the news domain hyperparameters are provided in Appendix B, Table B1.

4.1.2 GPT-4o

We fine-tune the GPT-4o model on a subset of the training set of each language pair for one epoch. The batch size and learning rate multiplier are automatically selected by the OpenAI fine-tuning API.[f] We fine-tune the GPT model for all language directions in the literary domain. After observing that the translation quality does not improve in the literary domain and due to high cost of fine-tuning, we discontinued this approach and did not fine-tune the GPT-4o model in the news domain.

4.1.3 LLaMa

We QLoRA (Dettmers *et al.* 2023) fine-tune the LLaMa 3.1 Instruct model with 8B parameters[g] using the *Unsloth*[h] library. We train the 4-bit quantized version of the LLaMa 3.1 model separately for each translation direction for one epoch, where we use the entire training set for each language pair. For datasets with less than 150k training sentences, we train for three epochs. The

[d] https://github.com/zeynepyirmibes/diversity_literalness_mt
[e] https://github.com/Helsinki-NLP/Opus-MT
[f] https://platform.openai.com/finetune
[g] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[h] https://github.com/unslothai/unsloth

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

self-attention layers are targeted during fine-tuning, which totals to only 1.05% of the parameters. We use a batch size of *64* and a learning rate of 2×10^{-4}.

4.2 Prompting

The major advantage of LLMs, especially instruct fine-tuned models, is the ability to adapt to different tasks without task-specific training. Based on this, we aim to observe the translation capabilities of the LLaMa and GPT-4o models in zero-shot and few-shot scenarios.

4.2.1 Zero-shot

In the zero-shot setting, we direct the LLaMa 3.1 Instruct and GPT-4o models to translate the given source text using the prompt shown below. We provide the name of the domain (*"literary"* or *"news"*) within the prompt, as it has been demonstrated that even a minimal cue like this can enhance the quality of domain-specific translations (Peng *et al.* 2023).

> Translate this {Domain} text from {Source Language} to {Target Language}. Only provide the translation, don't give any explanations.

4.2.2 Few-shot

We also investigate in-context learning through a few-shot setting, where we employ relevance-based sampling (Xu *et al.* 2020) to obtain translation examples for each language pair from its training set. For each test sample, we make use of the *mxbai-embed-large-v1*[i] English sentence embedding model to retrieve the most similar 10 sentences from the training set. We experiment with five different *n*-shot settings ($n = 2, 3, 5, 7, 10$) with the LLaMa 3.1 Instruct and GPT-4o models. For the news domain, we only use 2-shot and 7-shot prompting with GPT-4o. After caching the 10 relevance-sampled examples, we fetch the top *n* samples and plug them within the prompt. A 2-shot German-English prompt example is provided below.

> You are going to translate a sentence from German to English. Here are some example translations:
> German: Aber da sind sie, die dich verfolgen.
> English: But they are pursuing you.
>
> German: Man verfolgt dich!
> English: They are in pursuit of you!
>
> Now translate the following German sentence. Only provide the translation, don't give any explanations.
> German: "Du siehst wohl, daß sie dich verfolgen, und daß ich nichts Unwahres behaupte."
> English:

4.3 Decoding strategies

To ensure a meaningful analysis of lexical diversity and translation literalness, we prioritized maximizing the baseline translation quality, as metrics of stylistic variation are only interpretable when

[i] https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

the output is semantically and grammatically robust. Consequently, we utilized the standard recommended decoding configurations for each model family, which remained constant across all domains and language pairs. For NMT (OPUS-MT), we employed beam search with a width of 5, a standard practice for optimizing translation fidelity (Freitag and Al-Onaizan, 2017). For LLaMa-3.1, we used nucleus sampling (Holtzman *et al.* 2020) with *top-p* = 0.9 and *temperature* = 0.6 to balance fluency with creative expression. To account for the stochastic nature of sampling-based decoding, we performed three independent inference passes for each LLaMa 3.1 configuration (zero-shot, few-shot, and fine-tuned). The scores reported throughout this study represent the averages of these runs, ensuring that our evaluation is robust to the variance inherent in nondeterministic sampling. For GPT-4o, we fixed the decoding hyperparameters to the vendor-specified defaults (equivalent to *temperature* = 1.0, *top-p* = 1.0). By adhering to these well-established standards rather than over-engineering parameters for specific metrics, we ensured that our diversity results reflect the inherent behavior of these systems in high-quality translation regimes rather than artifacts of hyperparameter tuning.

5. Experiments

5.1 Datasets

In this work, we assess translation quality, lexical diversity, and translation literalness in MT outputs for the literary and news domains. The literary domain is chosen for its distinctive characteristics that strongly reflect these dimensions, featuring a rich vocabulary, varied sentence structures, and figurative language, with high lexical diversity and stylistic variation. Literary texts often feature nuanced meanings, metaphors, and cultural references that cannot always be translated literally unlike other domains (e.g. news or scientific text). The news domain, in contrast, typically employs more standardized vocabulary, factual reporting, and less stylistic variation, making it valuable for comparison. Including both domains allows us to examine MT performance on text types that differ substantially in linguistic complexity, style, and translatability.

We deduplicate the datasets for all language pairs in both domains and standardize the punctuation with a manually extracted punctuation dictionary. We tokenize Chinese text with the Jieba tokenizer[j] and Japanese text with the Mecab tokenizer for calculating the BLEU, lexical diversity, and translation literalness metrics.

5.1.1 Literary

We hold experiments with seven languages in the literary domain: Turkish, German, Spanish, French, Italian, Russian, and Chinese coupled with English. For the Turkish-English language pair, we use a dataset consisting of 93 manually aligned literary books (Yirmibeşoğlu *et al.* 2023).[k] We randomly select 90 books to construct the train and validation sets, and use the remaining three books as the test set. For the German, Spanish, French,[l] Italian, and Russian languages paired with English, we use the *Books* corpus which is a part of the OPUS collection (Tiedemann 2012). This is a multilingually aligned corpus consisting of 64 bitexts for 16 languages, with different number of parallel sentences for each language pair. We exploit the GuoFeng Webnovel Corpus v1 provided in the Eighth Conference on Machine Translation (WMT23) (Kocmi *et al.* 2023) for Chinese-English literary translation.

[j] https://github.com/fxsjy/jieba

[k] We use an extended version of the dataset introduced in the cited paper. Due to copyright issues, the dataset was requested from the authors and cannot be publicly shared.

[l] We manually fixed the two books that were in the wrong direction for the French-English language pair in the *Books* corpus.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table 1. Number of sentences in literary translation corpora

	DE-EN	ES-EN	FR-EN	IT-EN	RU-EN	TR-EN	ZH-EN
Train	49,855	91,858	125,473	30,720	15,884	503,026	1,893,749
Train (GPT)	5,158	5,289	5,321	4,937	5,613	6,068	5,606
Validation	514	515	520	521	517	4,075	695
Test	1,060	1,057	1,065	1,053	1,057	1,082	1,471
Total	51,429	93,430	127,058	32,294	17,458	508,183	1,895,915

Table 2. Number of sentences in news translation corpora

	DE-EN	ES-EN	FR-EN	HI-EN	IT-EN	JA-EN	RU-EN	TR-EN	ZH-EN
Train	435,776	498,495	405,200	4,923	102,218	8,496	375,989	206,066	441,315
Validation	529	530	525	520	528	500	528	518	528
Test	1,002	3,000	3,003	2,507	2,525	1,000	1,002	3,000	1,002
Total	437,307	502,025	408,728	7,950	105,271	9,996	377,519	209,584	442,845

We present the number of sentences in the literary corpora in Table 1. For fine-tuning in each language pair, we use the entire training set for the NMT and LLaMa models, while we only use a random portion of the sentences that correspond to 500k tokens including the prompt for the GPT-4o model (second row in the table) to reduce cost.

5.1.2 News

We experiment with nine languages in the news domain: Turkish, German, Spanish, French, Italian, Russian, Chinese, Japanese, and Hindi coupled with English. We use the SETIMES dataset for the Turkish-English pair, and the News Commentary (v18.1) dataset provided in the WMT24 training corpora for the rest of the eight languages. We use the latest WMT news test sets for testing (Appendix C, Table C1). The dataset statistics are provided in Table 2.

5.2 Results

In this section, we present the results of the experiments and make a comprehensive analysis of the models, model variations, domains, and languages. We first compare the models in terms of translation quality using four metrics used in the MT domain. Based on these results, we identify the best performing models. We continue with the analysis of the lexical diversity and translation literalness dimensions using the best models. This is followed by an error analysis of all the models using an error identification model. We then provide a detailed correlation analysis between the quality, diversity, and literalness metrics. To facilitate a structured overview of our empirical evaluation, each analytical subsection—translation quality, lexical diversity, translation literalness, error analysis, and correlation analysis—includes highlighted "Key Finding" boxes. These central takeaways are followed by detailed paragraphs grounding the findings.

While this study provides comprehensive insights, several limitations should be acknowledged. First, the fine-tuning of the GPT-4o model was constrained to approximately 500k tokens, due to

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table 3. Translation quality scores in the literary domain averaged over English→X and X→English language pairs. Models are grouped by family, with changes computed relative to the corresponding baseline (pretrained for NMT, zero-shot for LLMs). **Dark green** and **light green** indicate better (higher) scores; **dark red** and **light red** indicate worse (lower) scores. Light green/red denote a 0–10% change, whereas dark green/red denote a change exceeding 10%. Next to the average scores, the standard deviations are provided. Among systems, the highest score in each column is shown in **bold**. Human evaluation scores are additionally reported in *italic* if they are higher than all system scores

	BLEU		BLEURT		COMET		COMET-Kiwi	
Model	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN
Human	*100.00 ± 0.00*	*100.00 ± 0.00*	*0.976 ± 0.026*	*0.979 ± 0.002*	*0.952 ± 0.017*	*0.937 ± 0.008*	0.729 ± 0.025	0.719 ± 0.022
NMT	16.30 ± 5.85	19.34 ± 4.84	0.583 ± 0.040	0.599 ± 0.045	0.752 ± 0.027	0.735 ± 0.038	0.793 ± 0.020	0.771 ± 0.045
NMT + FT	22.43 ± 6.48	25.33 ± 4.69	0.620 ± 0.031	0.636 ± 0.028	0.777 ± 0.027	0.765 ± 0.019	0.788 ± 0.013	0.775 ± 0.022
GPT-4o	18.22 ± 5.90	21.82 ± 3.27	0.633 ± 0.031	0.646 ± 0.023	0.794 ± 0.029	0.779 ± 0.015	**0.824 ± 0.008**	**0.809 ± 0.016**
GPT-4o + 2-shot	20.28 ± 6.32	24.71 ± 3.84	0.645 ± 0.032	0.662 ± 0.022	0.797 ± 0.030	0.786 ± 0.017	0.818 ± 0.008	0.802 ± 0.016
GPT-4o + 3-shot	20.70 ± 6.22	25.19 ± 3.85	0.646 ± 0.032	0.667 ± 0.023	0.799 ± 0.031	0.788 ± 0.017	0.818 ± 0.009	0.801 ± 0.016
GPT-4o + 5-shot	21.10 ± 6.36	25.75 ± 4.01	0.647 ± 0.031	0.669 ± 0.023	**0.801 ± 0.030**	0.789 ± 0.017	0.818 ± 0.009	0.800 ± 0.016
GPT-4o + 7-shot	21.28 ± 6.32	26.07 ± 4.12	0.648 ± 0.031	0.670 ± 0.023	**0.801 ± 0.030**	**0.790 ± 0.017**	0.817 ± 0.008	0.799 ± 0.016
GPT-4o + 10-shot	21.50 ± 6.65	26.30 ± 4.13	**0.650 ± 0.030**	**0.671 ± 0.023**	0.800 ± 0.030	**0.790 ± 0.015**	0.817 ± 0.008	0.798 ± 0.016
GPT-4o + FT	14.18 ± 6.44	21.20 ± 4.32	0.517 ± 0.022	0.575 ± 0.048	0.705 ± 0.027	0.715 ± 0.038	0.670 ± 0.043	0.675 ± 0.053
LLaMa-3.1	13.31 ± 6.11	18.77 ± 4.21	0.586 ± 0.031	0.610 ± 0.024	0.762 ± 0.028	0.754 ± 0.014	0.797 ± 0.021	0.789 ± 0.020
LLaMa-3.1 + 2-shot	15.58 ± 6.11	21.84 ± 4.24	0.598 ± 0.032	0.641 ± 0.024	0.765 ± 0.029	0.771 ± 0.014	0.791 ± 0.018	0.791 ± 0.017
LLaMa-3.1 + 3-shot	15.95 ± 6.22	22.37 ± 4.27	0.602 ± 0.032	0.644 ± 0.024	0.769 ± 0.029	0.773 ± 0.014	0.791 ± 0.017	0.791 ± 0.017
LLaMa-3.1 + 5-shot	16.40 ± 6.20	22.84 ± 4.41	0.605 ± 0.032	0.647 ± 0.024	0.770 ± 0.028	0.775 ± 0.013	0.791 ± 0.016	0.789 ± 0.016
LLaMa-3.1 + 7-shot	16.67 ± 6.30	23.20 ± 4.33	0.606 ± 0.032	0.649 ± 0.024	0.771 ± 0.028	0.777 ± 0.014	0.789 ± 0.016	0.789 ± 0.017
LLaMa-3.1 + 10-shot	16.94 ± 6.30	23.57 ± 4.48	0.609 ± 0.032	0.651 ± 0.024	0.773 ± 0.029	0.778 ± 0.014	0.790 ± 0.015	0.787 ± 0.016
LLaMa3.1 + FT	**22.89 ± 6.63**	**27.36 ± 5.10**	0.626 ± 0.027	0.647 ± 0.026	0.783 ± 0.026	0.775 ± 0.014	0.777 ± 0.010	0.766 ± 0.013

cost limitations. We conducted an ablation study to evaluate the impact of this asymmetric training budget (Section 5.2.7). Second, there are notable dataset size imbalances across the different language pairs evaluated in this study (as detailed in Tables 1 and 2), which may affect the averaged results by disproportionately weighting high-resource languages. Therefore to support our key findings, we also provide a detailed statistical significance analysis in Appendix H (Tables H1 and H2). Finally, the literalness metrics employed in our evaluation (USW and NM) inherently depend on the quality of the underlying word alignments; any inaccuracies in the alignment process could introduce noise into the literalness scores.

5.2.1 Translation quality

Tables 3 and 4 show the BLEU, BLEURT, COMET, and COMET-Kiwi scores in the literary domain and the news domain, respectively, for all the models used in this work. We also evaluate the original (human) translation to analyze how models perform compared to the gold translation. The scores are given as the averages of all the languages used in that domain and depicted as EN → X (English-to-X) and X → EN (X-to-English). Although the average score for a model may not present an exact figure since it combines scores of different languages, it gives an overview about the performance of the model in that metric. The language-specific scores are visualized as bar plots for the BLEU metric (Figures 1 and 2) and the COMET metric (Figures 3 and 4). To ensure visual clarity, these plots represent only the highest-shot configuration for each LLM-based model family (e.g., 10-shot for GPT and 7-shot for LLaMa), rather than displaying all tested few-shot variations. The results for the other metrics are presented in Appendix D (Figures D1, D2, D3 and D4). Directional asymmetries observed for translation quality are commented in Appendix G.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table 4. Translation quality scores in the news domain averaged over English→X and X→English language pairs

	BLEU		BLEURT		COMET		COMET-Kiwi	
Model	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN
Human	*100.00 ± 0.00*	*100.00 ± 0.00*	*0.977 ± 0.014*	*0.980 ± 0.001*	*0.957 ± 0.011*	*0.950 ± 0.007*	0.833 ± 0.018	0.818 ± 0.015
NMT	25.31 ± 10.43	26.89 ± 9.44	0.642 ± 0.120	0.686 ± 0.070	0.799 ± 0.087	0.820 ± 0.046	0.792 ± 0.076	0.809 ± 0.046
NMT + FT	27.08 ± 9.68	26.95 ± 8.70	0.671 ± 0.091	0.688 ± 0.067	0.816 ± 0.078	0.821 ± 0.046	0.809 ± 0.064	0.808 ± 0.047
GPT-4o	**30.46 ± 6.74**	32.81 ± 5.63	0.734 ± 0.043	0.752 ± 0.016	**0.869 ± 0.025**	0.868 ± 0.014	**0.857 ± 0.014**	**0.846 ± 0.008**
GPT-4o + 2-shot	29.87 ± 6.60	32.70 ± 5.25	0.734 ± 0.044	0.755 ± 0.016	**0.869 ± 0.026**	**0.869 ± 0.014**	**0.857 ± 0.014**	**0.846 ± 0.009**
GPT-4o + 7-shot	29.96 ± 6.76	**32.90 ± 5.00**	**0.735 ± 0.044**	**0.756 ± 0.015**	**0.869 ± 0.026**	**0.869 ± 0.014**	**0.857 ± 0.014**	**0.846 ± 0.008**
LLaMa-3.1	21.75 ± 8.34	28.13 ± 6.07	0.679 ± 0.046	0.718 ± 0.023	0.834 ± 0.027	0.846 ± 0.012	0.828 ± 0.020	0.833 ± 0.010
LLaMa-3.1 + 2-shot	22.95 ± 8.02	29.28 ± 5.68	0.685 ± 0.044	0.731 ± 0.018	0.837 ± 0.027	0.854 ± 0.012	0.831 ± 0.021	0.835 ± 0.010
LLaMa-3.1 + 3-shot	22.93 ± 8.12	29.50 ± 5.58	0.686 ± 0.044	0.731 ± 0.018	0.837 ± 0.028	0.855 ± 0.012	0.831 ± 0.021	0.835 ± 0.010
LLaMa-3.1 + 5-shot	23.10 ± 8.08	29.61 ± 5.54	0.688 ± 0.044	0.732 ± 0.018	0.839 ± 0.027	0.855 ± 0.012	0.832 ± 0.020	0.835 ± 0.010
LLaMa-3.1 + 7-shot	23.07 ± 8.10	29.70 ± 5.48	0.689 ± 0.044	0.732 ± 0.018	0.839 ± 0.027	0.855 ± 0.012	0.832 ± 0.020	0.834 ± 0.010
LLaMa-3.1 + 10-shot	23.09 ± 8.08	29.84 ± 5.50	0.689 ± 0.044	0.732 ± 0.018	0.840 ± 0.027	0.855 ± 0.011	0.832 ± 0.020	0.834 ± 0.010
LLaMa3.1 + FT	26.77 ± 8.13	30.65 ± 4.88	0.702 ± 0.054	0.727 ± 0.019	0.848 ± 0.038	0.851 ± 0.013	0.835 ± 0.024	0.828 ± 0.014

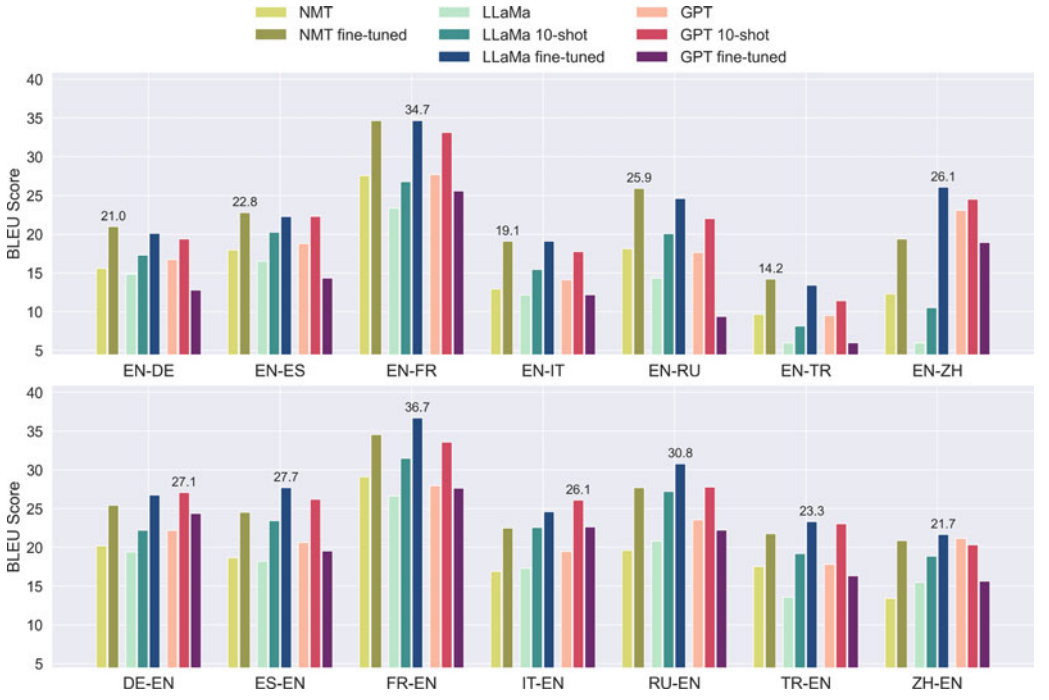


Figure 1. BLEU scores of the models for English→X (above) and X→English (below) language pairs for the literary domain.

Key finding: Metric divergence

While fine-tuned open-source models (NMT and LLaMa) frequently lead in BLEU for the literary domain, GPT-4o remains the most robust performer across neural metrics, outperforming open-source models in both domains.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

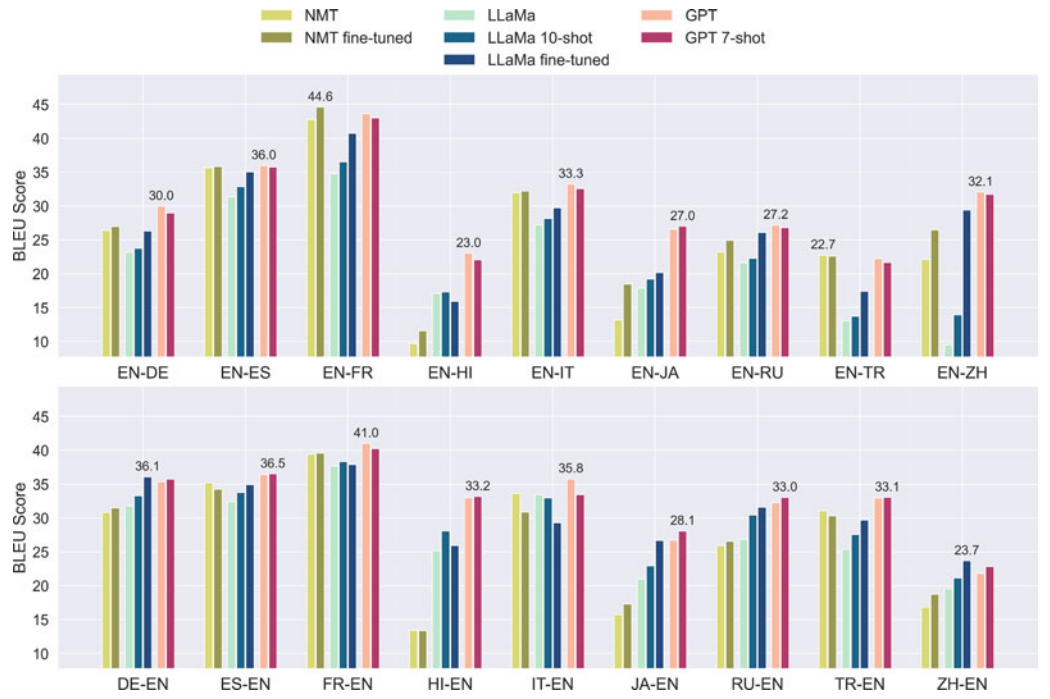


Figure 2. BLEU scores of the models for English→X (above) and X→English (below) language pairs for the news domain.

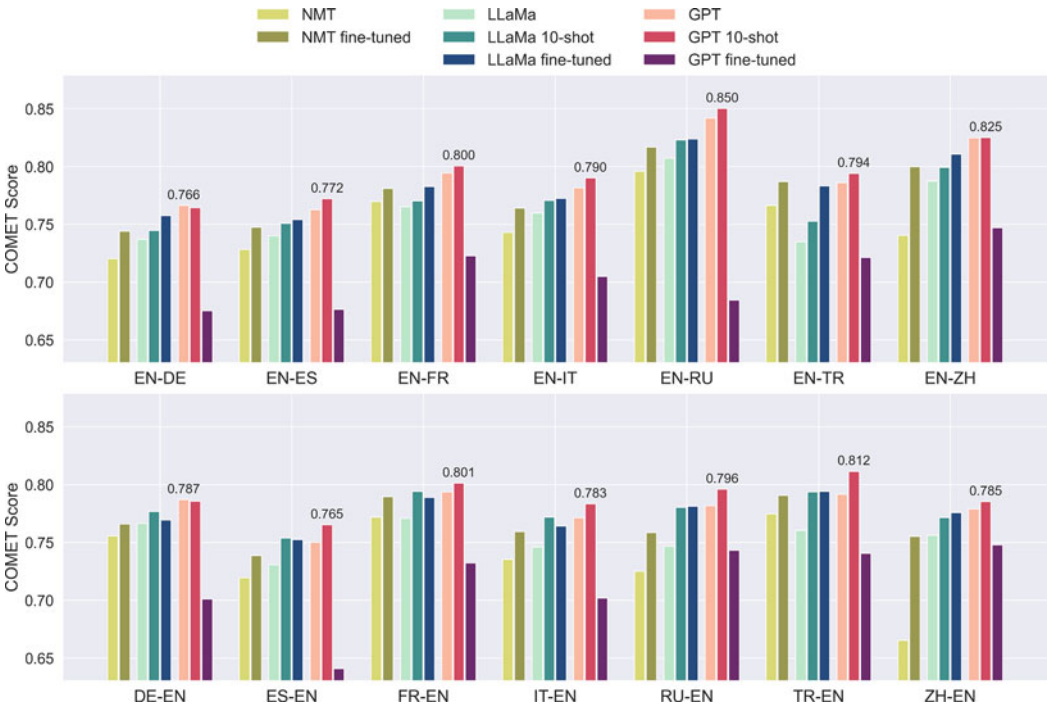


Figure 3. COMET scores of the models for English→X (above) and X→English (below) language pairs for the literary domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

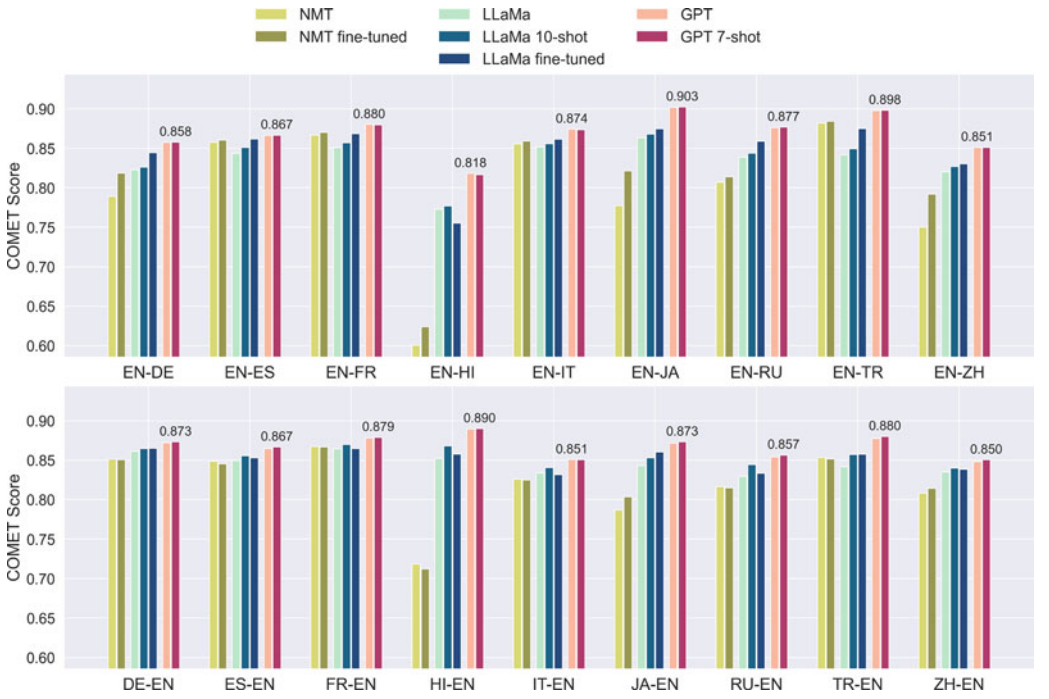


Figure 4. COMET scores of the models for English→X (above) and X→English (below) language pairs for the news domain.

Our analysis reveals a divergence between metric narratives. In the literary domain, BLEU (Figure 1) favors fine-tuned models: LLaMa 3.1 leads in most X→EN directions, while NMT leads in EN→X. COMET results (Figure 3) shift the lead to GPT-4o, which achieves the highest scores in both directions with few-shot prompting. For the news domain (Figures 2 and 4), GPT dominance is more uniform across metrics, leading in the majority of language directions and maintaining high quality in Hindi and Japanese where NMT and LLaMa show significant weaknesses. This finding shows the strength of the GPT model in a more generic domain (e.g. news) and displays the redundancy of resorting to fine-tuning for GPT in such domains where pre-training data is abundant. Notably, while fine-tuning benefits NMT and LLaMa in the literary domain—particularly for EN→X directions—it consistently degrades GPT performance, likely due to catastrophic forgetting or data constraints. Across all configurations and metrics, human translations remain the upper bound, consistently exceeding MT systems in COMET and BLEURT scores.

Key finding: Reference-free metric discrepancy

While BLEURT and COMET reinforce GPT-4o's dominance and the clear superiority of human translations, COMET-Kiwi (reference-free) frequently ranks MT systems above human translations in the literary domain, likely overestimating surface-level fluency at the expense of stylistic creativity.

BLEURT results closely correlate with COMET trends, identifying GPT-4o 10-shot (literary) and 7-shot (news) as the top-performing systems, with few-shot gains appearing more pronounced in BLEURT than in COMET. Crucially, human translations consistently maintain a significant lead over all MT systems in BLEURT across both domains. Conversely, the

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

nonreference-based COMET-Kiwi produces unexpected patterns: in the literary domain, GPT-4o zero-shot achieves the highest scores, while prompting offers minimal benefit. Most notably, nearly all MT systems—except the fine-tuned GPT-4o—surpass human scores in this metric for literary texts, a trend we report in more detail in Appendix D. While GPT leads again in the news domain and benefits slightly from few-shot prompting, the systematic dominance of MT scores over human translations in literary contexts suggests a limitation in reference-free quality estimation. By prioritizing surface-level coherence and fluency, COMET-Kiwi may fail to capture the stylistic nuances and the creative "free style" intention in human literary translation, leading it to reward MT's predictable adequacy more generously.

Key finding: Fine-tuning divergence

Fine-tuning is transformative for open-source models (NMT and LLaMa) in the literary domain but causes a performance collapse in GPT-4o, suggesting that for proprietary LLMs, few-shot prompting is a more robust adaptation strategy than small-scale gradient updates.

In the literary domain, fine-tuning on in-domain parallel data acts as a primary performance driver for NMT and LLaMa. The NMT baseline improves on average by ≈6 BLEU in both directions, and fine-tuned LLaMa 3.1 achieves its peak performance after fine-tuning (22.89 EN→X, 27.36 X→EN), even surpassing various GPT-4o configurations. However, this trend does not hold for GPT-4o, where average performance drops from 18.22 to 14.18 BLEU for EN→X and from 21.82 to 21.20 for X→EN. This can suggest catastrophic forgetting or insufficient/biased fine-tuning data (only 500k tokens). Assuming this could result from underfitting, we conducted further experiments with different batch sizes and number of epochs, but to no avail. Training with a larger set was also not a viable option due to very high fine-tuning costs. We conclude that few-shot prompting is much cheaper and much more effective with respect to fine-tuning in terms of translation quality for GPT, a deduction confirmed in all language pairs. Observing the same trend in the news domain for GPT fine-tuning, we discontinue this method in the news domain. In the news domain, the impact of FT is more localized; it significantly improves the EN→X direction for LLaMa, while providing modest, though not aggregate-significant, improvements for NMT. For NMT, X→EN gains in news also remain negligible.

Key finding: The English-target ceiling

English-centric bias creates a performance ceiling for $X \rightarrow EN$, leaving $EN \rightarrow X$ as the primary beneficiary of fine-tuning due to its massive improvement headroom.

Our results also highlight a systematic difference between directions, with X→EN often scoring higher than EN→X. All MT settings, we experiment with perform much better for the English target language than for X target languages, even when evaluated on the same parallel test set. This discrepancy likely stems from the "English-centric" bias inherent in the massive pre-training corpora of LLMs and the extensive English-side data available for NMT baselines. Consequently, because $EN \rightarrow X$ performance starts from a lower baseline, it presents a significantly larger improvement headroom. This explains why adaptation methods like fine-tuning yield such substantial gains for $EN \rightarrow X$—effectively bridging the resource gap—while $X \rightarrow EN$ remains relatively stable, especially in the news domain, due to its already saturated near-peak performance.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Key finding: Few-shot efficiency and saturation

Few-shot prompting yields consistent quality gains in the literary domain, whereas news domain show saturation, with GPT-4o zero-shot performance already approaching the ceiling for neural metrics.

Few-shot prompting serves as a highly efficient and cost-effective alternative to fine-tuning, particularly for GPT-4o and LLaMa in the literary domain. For GPT-4o, increasing the context from zero-shot to 10-shot results in a steady rise from 20.02 (18.22 EN→X and 21.82 X→EN) to 23.90 (21.50 EN→X and 26.30 X→EN) average BLEU, with the most significant gains occurring within the first 5 shots. LLaMa 3.1 follows a similar trajectory, rising from 13.31 to 16.94 BLEU for EN→X and from 18.77 to 23.57 BLEU for X→EN. Conversely, the news domain exhibits a "saturation effect." For GPT-4o, few-shot prompting provides no systematic benefit and occasionally results in minor BLEU declines, likely because the model's pre-training data is already abundant in news-like corpora. While LLaMa 3.1 still shows modest improvements from few-shot prompting in the news domain, it never reaches the zero-shot baseline of GPT-4o. Fine-tuning remains the clear lever for LLaMa, especially for EN→X language pairs. These results suggest that while prompting is essential for bridging the gap in specialized domains like literature, its utility is limited in generic domains where zero-shot performance is already near its peak.

Key finding: GPT-4o and the prompting paradigm

GPT-4o establishes the performance ceiling among MT systems under neural quality metrics across both domains.

Our results suggest that GPT-4o consistently delivers the highest translation quality under neural quality metrics, with its zero- and few-shot configurations leading in both news and literary domains. While the BLEU metric occasionally highlights fine-tuned NMT and LLaMa models in the literary domain, neural metrics—COMET, BLEURT, and COMET-Kiwi—consistently favor GPT-4o. This underscores the model's superior cross-domain generalization and robustness. Under neural quality metrics, the overall hierarchy is broadly $GPT \geq LLaMa \geq NMT$, and few-shot prompting is an efficient quality driver mainly in the literary domain, while news shows saturation for GPT-4o (Appendix H).

5.2.2 Best performing models

Before moving on to analyzing lexical diversity and translation literalness, we first identify the models that show high performance in translation quality. The motivation behind this is to compare models with similar and high translation quality rather than comparing all models, as suggested in Raunak *et al.* (2023). Without this consideration high lexical diversity and translation nonliteralness might be misleading, as poor translation performance can also result in apparent lexical richness. A sentence could appear to be lexically diverse simply due to inaccuracies in translation.

To evaluate diversity and literalness, we compared top-performing configurations against their pretrained counterparts. We identified the "best" models by aggregating rankings across four quality metrics (BLEU, COMET, BLEURT, and COMET-Kiwi) for all language pairs; the models with the highest cumulative ranking scores were selected. For the literary domain, the top models were GPT (10-shot, base, 7-shot) and fine-tuned LLaMA and NMT. In the news domain, the top performers were GPT (7-shot, base, 2-shot) and the same fine-tuned models. Our final analysis includes these top configurations alongside their pretrained counterparts. We complete the list by

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table 5. Lexical diversity (TTR, Yule's I, and MTLD) and translation literalness (USW and NM) scores in the literary domain averaged over English→X and X→English language pairs. Higher the literalness metrics, the less literal the translation

	TTR		Yule's I		MTLD		USW		NM	
Model	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN
Human	*0.289 ± 0.104*	0.190 ± 0.034	*24.2 ± 39.0*	4.3 ± 1.5	*369.4 ± 507.8*	120.1 ± 14.9	*30.4 ± 6.4*	*28.7 ± 4.9*	*17.9 ± 3.8*	*17.9 ± 3.8*
NMT	0.273 ± 0.091	0.173 ± 0.027	14.5 ± 15.8	3.1 ± 1.1	256.1 ± 227.5	102.7 ± 19.4	19.6 ± 8.5	19.8 ± 10.0	14.2 ± 6.5	14.8 ± 6.7
NMT + FT	0.275 ± 0.097	0.171 ± 0.031	17.4 ± 24.7	3.1 ± 1.1	273.8 ± 295.3	102.7 ± 11.1	23.1 ± 8.2	21.6 ± 6.9	16.1 ± 7.4	15.6 ± 6.9
GPT-4o	0.274 ± 0.097	0.184 ± 0.025	16.5 ± 19.8	4.0 ± 1.1	306.6 ± 327.5	125.9 ± 12.9	19.1 ± 8.7	20.8 ± 6.3	14.1 ± 6.5	15.0 ± 5.4
GPT-4o + 10-shot	**0.285 ± 0.098**	**0.192 ± 0.024**	**18.4 ± 23.6**	**4.4 ± 1.1**	**322.4 ± 362.6**	**130.2 ± 11.8**	21.0 ± 8.5	21.1 ± 6.5	15.0 ± 6.3	15.0 ± 5.6
LLaMa-3.1	0.273 ± 0.089	0.184 ± 0.028	14.4 ± 16.0	4.0 ± 1.5	202.4 ± 124.1	119.2 ± 15.0	18.8 ± 9.2	16.8 ± 7.1	13.6 ± 6.5	13.6 ± 5.8
LLaMa-3.1 + 10-shot	0.275 ± 0.093	0.179 ± 0.027	15.6 ± 19.6	3.6 ± 1.1	192.9 ± 114.1	114.9 ± 9.5	20.7 ± 9.2	19.6 ± 7.4	14.9 ± 6.3	15.4 ± 5.9
LLaMa3.1 + FT	0.264 ± 0.085	0.170 ± 0.031	15.2 ± 21.7	3.0 ± 1.0	119.9 ± 54.7	96.6 ± 9.9	**27.0 ± 7.5**	**25.6 ± 5.9**	**16.7 ± 7.1**	**16.8 ± 6.4**

Table 6. Lexical diversity (TTR, Yule's I, and MTLD) and translation literalness (USW and NM) scores in the news domain averaged over English→X and X→English language pairs

	TTR		Yule's I		MTLD		USW		NM	
Model	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN
Human	*0.227 ± 0.088*	0.180 ± 0.033	*14.1 ± 19.4*	*3.7 ± 1.7*	172.4 ± 134.6	99.9 ± 5.9	*25.4 ± 8.9*	27.7 ± 11.9	*18.2 ± 8.6*	18.1 ± 8.5
NMT	**0.224 ± 0.087**	0.163 ± 0.023	11.7 ± 13.6	2.3 ± 0.7	167.6 ± 126.6	85.5 ± 7.7	23.5 ± 13.0	27.4 ± 13.8	16.1 ± 9.7	17.1 ± 8.4
NMT + FT	0.219 ± 0.091	0.170 ± 0.028	11.5 ± 14.1	2.8 ± 1.1	171.1 ± 132.5	90.9 ± 11.6	22.6 ± 11.9	29.2 ± 12.6	16.2 ± 8.8	16.3 ± 8.1
GPT-4o	0.221 ± 0.091	0.171 ± 0.029	12.2 ± 15.8	2.9 ± 1.1	177.7 ± 142.0	96.3 ± 3.0	21.4 ± 10.6	26.1 ± 12.4	15.6 ± 8.8	15.1 ± 8.1
GPT-4o + 7-shot	**0.224 ± 0.090**	**0.181 ± 0.031**	**12.7 ± 16.5**	**3.5 ± 1.3**	**184.3 ± 150.2**	**104.2 ± 3.0**	21.8 ± 10.3	28.1 ± 11.8	16.2 ± 9.0	15.8 ± 8.0
LLaMa-3.1	**0.224 ± 0.090**	0.177 ± 0.029	11.8 ± 14.4	3.2 ± 1.1	145.6 ± 87.1	96.0 ± 4.1	22.2 ± 11.2	25.4 ± 12.2	15.8 ± 8.4	**18.5 ± 14.5**
LLaMa-3.1 + 10-shot	0.221 ± 0.089	0.174 ± 0.029	11.7 ± 14.9	3.1 ± 1.2	148.3 ± 97.3	95.8 ± 3.6	22.4 ± 11.3	28.0 ± 12.8	16.1 ± 8.9	16.4 ± 8.4
LLaMa3.1 + FT	0.213 ± 0.085	0.176 ± 0.028	10.1 ± 12.7	3.1 ± 1.1	145.0 ± 101.5	88.0 ± 10.6	**23.8 ± 11.2**	**31.3 ± 11.3**	**16.6 ± 8.8**	16.9 ± 7.7

also including LLaMa 10-shot as the top few-shot version of LLaMa. This selection provides a representative group of pretrained, few-shot, and fine-tuned models across three families, excluding only the low-performing GPT fine-tuned model.

Tables 5 and 6 depict scores of the selected models for the lexical diversity metrics (TTR, Yule's I, MTLD) and the translation literalness metrics (USW, NM) averaged over EN→X and X→EN language pairs in, respectively, the literary domain and the news domain. In Sections 5.2.3 and 5.2.4, we analyze the diversity and literalness results by referring to both the averages in these tables and the language-specific scores provided in the figures. The scores of all models across all metrics and language pairs are provided in our Github page.[m]

5.2.3 Lexical diversity

Among the lexical diversity metrics, MTLD is the most robust to text length. We thus present the MTLD results for the selected models in Figures 5 and 6. The lexical diversity scores in the English-Turkish direction are much higher than those in the other directions in the literary domain due to the agglutinative nature of Turkish. Thus, we present them in a separate scale within the literary domain bar plot. The results for the TTR and Yule's I metrics are provided in Appendix E (Figures E1, E2, E3 and E4) Directional asymmetries observed for lexical diversity are commented in Appendix G.

[m] https://github.com/zeynepyirmibes/diversity_literalness_mt

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

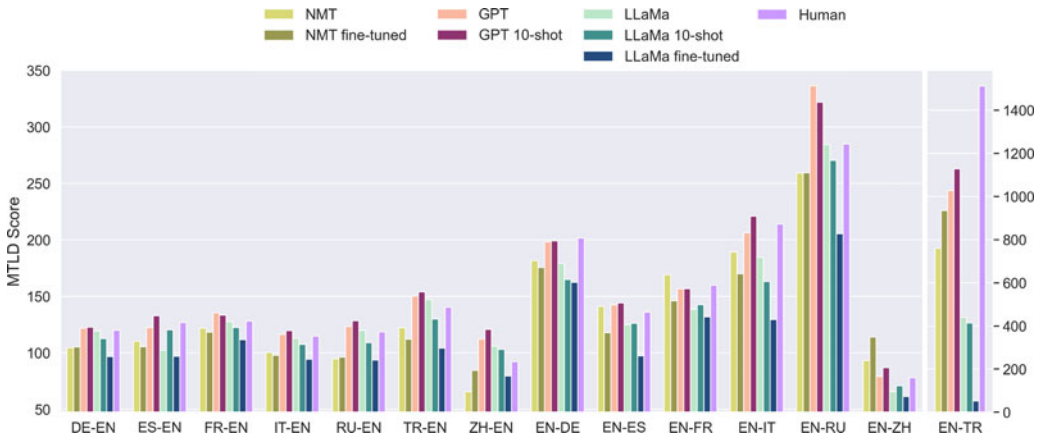


Figure 5. MTLD scores of pretrained and best-performing models for the literary domain.

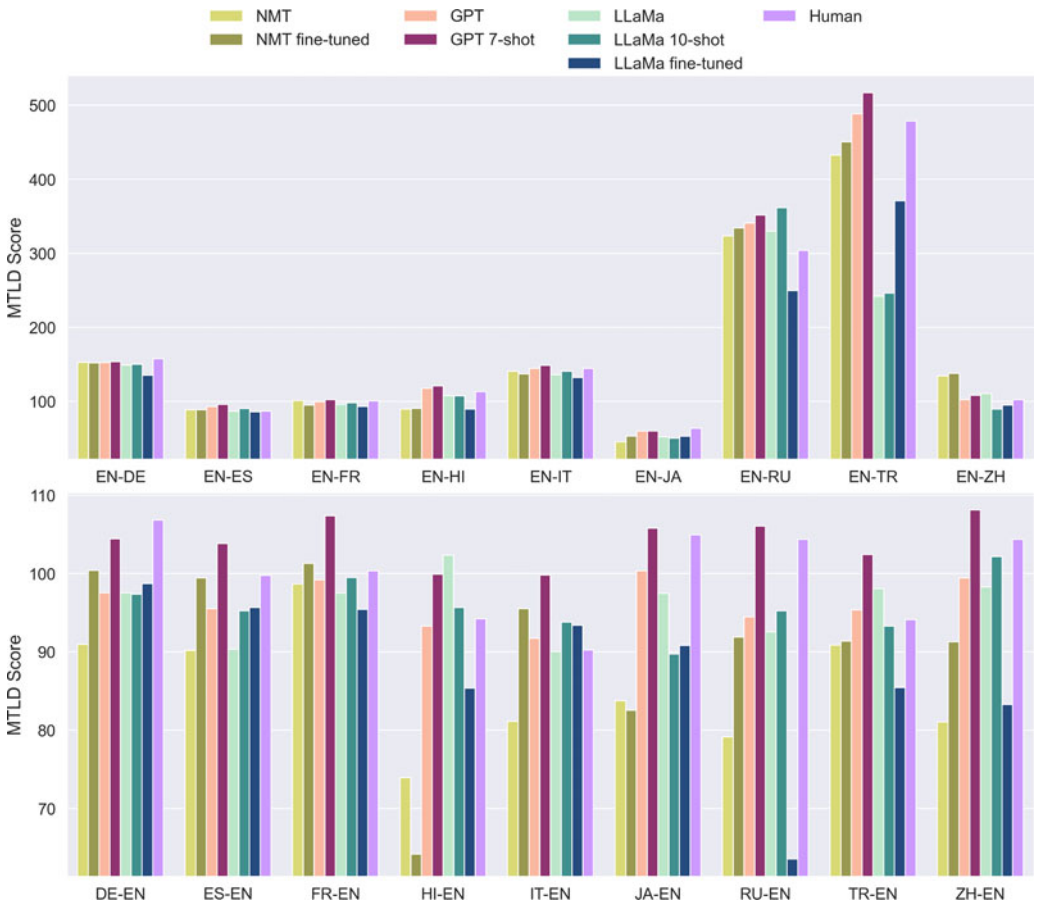


Figure 6. MTLD scores of pretrained and best-performing models for the news domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Key finding: GPT and the prompting advantage

Across both domains, GPT with few-shot prompting generally produces the most lexically diverse translations.

In the literary domain, GPT few-shot achieves the highest average MTLD (322.4 EN→X; 130.2 X→EN). In the news domain, GPT few-shot leads in average MTLD, Yule's I, and X→EN TTR, while EN→X TTR is tied with NMT and LLaMa. Few-shot prompting is associated with a measurable and statistically significant lexical-richness boost over zero-shot in the news domain, with an average MTLD increase of + 6.6 (EN→X) and + 7.9 (X→EN), though this increase is not aggregate-significant in the literary domain. GPT 7-shot is the top performer for most language directions in the news domain, except for specific cases such as EN→RU, EN→ZH, and HI→EN, where it is surpassed by NMT fine-tuned for EN→ZH and by LLaMa variants for the others. Unlike LLaMa, GPT's lexical diversity increases with few-shot prompting in the news domain, suggesting that prompting is associated with broader lexical variation in the news domain without reducing translation quality. However, this trend is nonsignificant in the literary domain (Appendix H).

Key finding: The LLaMa fine-tuning trade-off

Fine-tuning LLaMa tends to improve translation quality, but it reduces lexical diversity in the literary domain. Few-shot prompting is generally less damaging than fine-tuning, though its gains over zero-shot remain limited and inconsistent.

Although fine-tuning improves LLaMa's translation quality scores, it reduces lexical diversity in the literary domain, where fine-tuned LLaMa yields the lowest MTLD among all systems. In the news domain, LLaMa fine-tuning lowers the average lexical diversity scores across all metrics, most sharply in EN→X Yule's I and X→EN MTLD, though the pattern is not uniform across language directions. Overall, the results point to a trade-off for LLaMa: gains in BLEU and COMET are accompanied by narrower lexical variation, especially in the literary domain. Notably, the fine-tuned NMT model, despite being a smaller bilingual system, outperforms the fine-tuned LLaMa in all literary language pairs for MTLD, suggesting that multilingual LLM fine-tuning may constrain lexical richness more strongly. Compared with fine-tuning, few-shot prompting is generally less damaging for LLaMa, preserving more lexical richness across both domains. However, its gains over zero-shot remain limited and inconsistent: in the literary domain it shows only small EN→X improvements alongside X→EN declines, and in the news domain it offers little overall benefit, even though it remains more lexically rich than fine-tuned LLaMa in 12 of 18 language directions.

Key finding: System-level ranking across model families

Across both domains, GPT is generally the strongest model family for lexical diversity, LLaMa the weakest, and NMT occupies an intermediate position.

NMT shows a more balanced lexical diversity profile than LLaMa, although it still falls short of GPT. In the literary domain, pretrained NMT outperforms all LLaMa variants on average for MTLD in EN→X, while fine-tuned NMT ranks behind the GPT configurations but ahead of fine-tuned LLaMa. The NMT model displays the highest lexical diversity for English→French and English→Chinese in the literary domain. In the news domain, NMT again remains below GPT overall, yet fine-tuning brings modest improvements, especially in the X→EN direction; this contrasts with the literary domain, where fine-tuning more often reduces NMT lexical diversity across language directions. However, these changes in NMT's lexical diversity upon fine-tuning

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

are not aggregate-significant in either domain, supporting the conclusion that NMT is lexically stable under fine-tuning. Taken together, the system-level comparison places GPT as the strongest model family for lexical diversity, NMT as an intermediate and comparatively stable option, and LLaMa as the weakest, particularly after fine-tuning.

Key Finding: Machine *versus* human lexical richness

GPT few-shot can surpass human average lexical diversity scores.

Humans achieve the highest average TTR and Yule's I scores for EN→X across both domains, with especially high values in morphologically rich languages such as Turkish (EN-TR). However, in the X→EN direction, GPT 10-shot (130.2 MTLD) exceeds the human average (120.1 MTLD) in the literary domain, and GPT 7-shot exceeds the human average MTLD in both EN→X and X→EN in the news domain. This is in line with recent findings that newer LLMs can sometimes exceed human lexical variation, especially in translation to English (Yao and Fan 2025). Humans still employ a more nuanced vocabulary density when translating into morphologically rich languages like Turkish. Similarly for German, frequent use of compound words contributes to a larger inventory of surface word forms in human translation, especially in a challenging domain such as literature.

5.2.4 Translation literalness

We measure the translation literalness of the MT outputs using the unaligned source words (USW) and nonmonotonicity (NM) metrics. We present the NM results for the literary domain in Figure 7 and the USW results for the news domain in Figure 8. The literary USW and news NM plots are provided in Appendix F (Figures F1 and F2). High scores in these metrics correspond to more unaligned words and word order deviations between the translations, signaling a less literal translation.

Key finding: Adaptation shifts models away from literalness

Zero-shot LLMs and pretrained NMT produce relatively literal translations. Adaptation often shifts outputs toward less literal translations, with fine-tuned LLaMa showing the strongest shift.

In the literary domain (Table 5), pretrained NMT and GPT zero-shot yield very similar USW and NM values, establishing a relatively literal baseline. Few-shot prompting generally increases these metrics, and fine-tuning produces the largest shift: fine-tuned LLaMa has the highest average USW and NM values, while fine-tuned NMT also ranks above the GPT and 10-shot LLaMa variants on the table averages. A similar pattern appears in the news domain (Table 6), where fine-tuned LLaMa is the least literal system on the aggregate USW averages and is the least literal configuration in 14/18 language directions. Under USW, fine-tuned NMT is less literal than pretrained NMT in 12/18 directions, though this shift for NMT is only a tendency and not a statistically robust aggregate effect. GPT few-shot is less literal than zero-shot GPT in 17/18. Overall, these results suggest that adaptation, especially fine-tuning for LLaMa, pushes outputs toward freer word orders and less literal lexical choices.

Key finding: LLaMa shows the largest within-family shift

LLaMa appears the most responsive to adaptation: few-shot prompting already reduces literalness, and fine-tuning drives the largest shift away from source-close translation.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

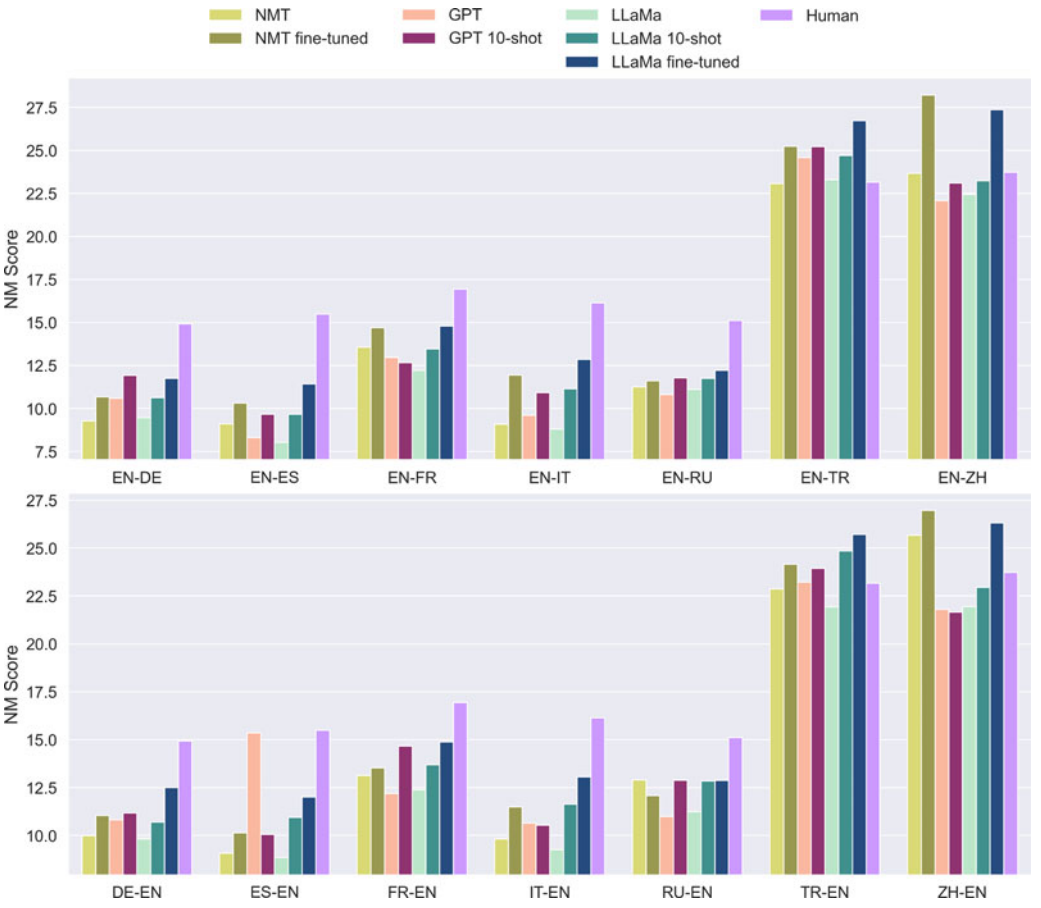


Figure 7. NM scores of pretrained and best-performing models in the literary domain.

Within families, NMT fine-tuning increases both USW and NM relative to its pretrained baseline in the literary domain, an effect most visible for EN→X, with RU-EN as the only exception. GPT becomes less literal under few-shot prompting in most news domain language directions, yet its within-family shifts remain smaller than those of LLaMa, whose few-shot and fine-tuning adaptations produce the steepest increases in USW and NM across the selected systems. In the literary domain, fine-tuned LLaMa most often yields the least literal outputs, although a small number of individual language pairs and metric comparisons favor fine-tuned NMT or GPT 10-shot instead.

Key finding: GPT trails fine-tuned models on nonliteralness

Unlike the lexical diversity picture, where GPT few-shot dominates, GPT-4o configurations remain comparatively literal on literalness metrics.

The average literalness results place fine-tuned LLaMa and to a lesser extent fine-tuned NMT above GPT configurations. The aggregate significance results (Appendix H) support this contrast most strongly for LLaMa. In the literary domain, GPT shows only modest literalness changes with few-shot prompting and does not match the nonliteralness of fine-tuned open-source models, except for EN-DE and RU-EN. In the news domain, GPT produces the most literal EN→X outputs under both USW and NM on average. This contrast suggests that lexical diversity and

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

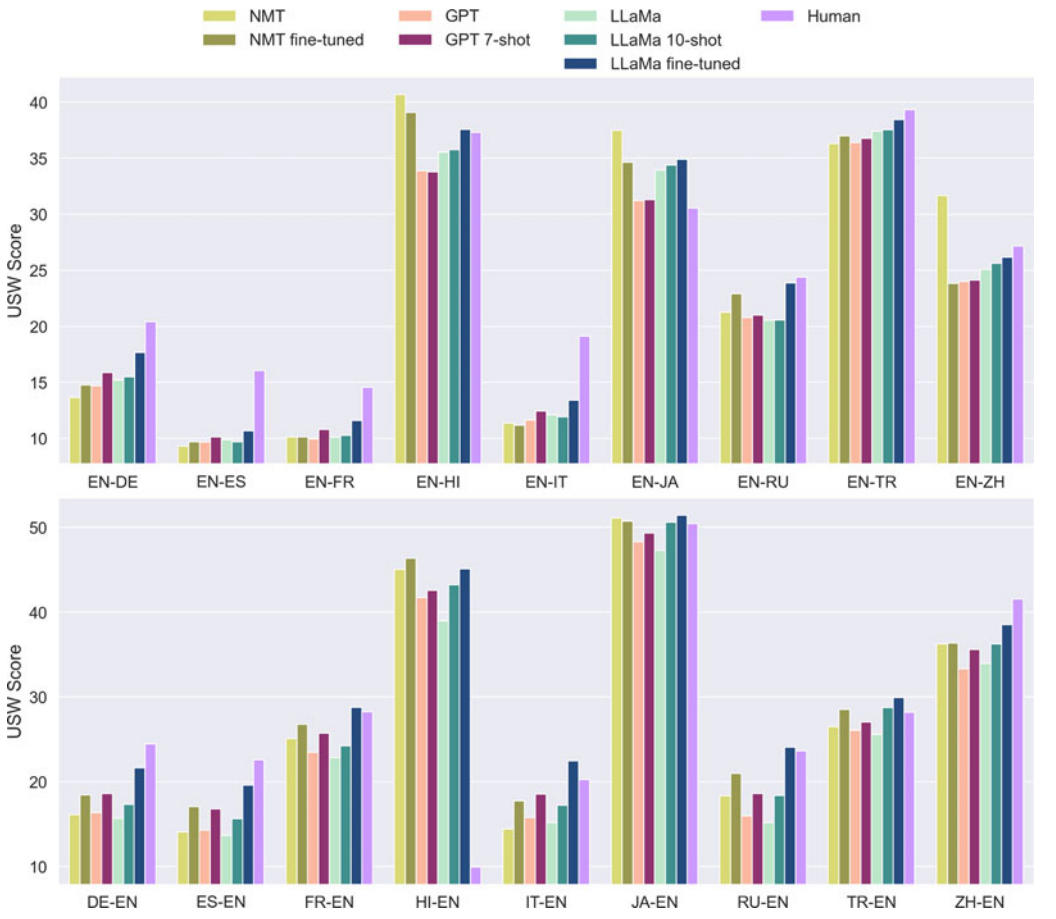


Figure 8. USW scores of pretrained and best-performing models in the news domain.

translation literalness capture partly distinct properties across model families: GPT tends to score higher on lexical-richness metrics, whereas fine-tuned LLaMa, and in some average comparisons fine-tuned NMT, deviate more from source alignment and word order.

Key finding: USW and NM are complementary

USW and NM capture complementary aspects of literalness. Both agree that adaptation reduces literalness, but they diverge on specific configurations, particularly in the news domain.

USW (alignment) and NM (word order) capture complementary aspects of literalness. USW measures how many source words remain unaligned in the translation, whereas NM measures how strongly the target word order departs from the source. A system can therefore stay relatively close to the source in alignment while still reordering it substantially, or vice versa. This pattern is visible in the news domain, where the two metrics diverge at the configuration level. Under X→EN USW, zero-shot LLaMa is among the more literal settings, whereas fine-tuned LLaMa is the least literal; under X→EN NM, zero-shot LLaMa is the most nonliteral setting among the MT systems. GPT is the most literal configuration family by NM and remains relatively literal by USW. These differences show why USW and NM should be treated as complementary rather than

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

interchangeable measures. Despite metric-specific divergences, both USW and NM point to the same broader tendency: adaptation often reduces literalness, but the magnitude and exact ranking depend on direction, model family, and metric.

Key finding: Humans translate less literally than most MT systems

Human translations are, on average, less literal than most MT configurations, though fine-tuned LLaMa and, in some settings, fine-tuned NMT surpass human nonliteralness in specific directions.

In the literary domain, both USW and NM indicate that human translations are less literal than the system averages overall, with exceptions for Turkish and Chinese (EN-TR, EN-ZH, TR-EN, and ZH-EN), where fine-tuned LLaMa or fine-tuned NMT can be less literal. In the news domain, human translations remain less literal than MT systems on average for EN→X, but fine-tuned NMT and LLaMa configurations exceed the human average in X→EN depending on the metric: fine-tuned NMT and LLaMa exceed the human average under USW, while zero-shot LLaMa is the least literal overall under NM.

5.2.5 Error analysis

We perform error analysis in both domains using the XCOMET-XXL evaluation model (Guerreiro *et al.* 2024). This model gives an overall system score and minor, major, and critical error annotations on translated sentences with a confidence score. We normalize the total number of translation errors for each error category with the number of sentences in the test set and come up with an error rate. We only take into account error annotations with larger than 50% confidence. We report the system scores and error rates averaged over all language directions in Figure 9.

Key finding: GPT few-shot sets the error-rate ceiling

Across both domains, GPT with few-shot prompting achieves the best balance of translation quality and low error rates, with critical error reductions plateauing beyond 5–7 shots.

In the literary domain, GPT-4o system scores steadily improve from zero-shot to 10-shot, with error rates—especially critical errors—decreasing substantially, while gains plateau beyond 5/7 shots; GPT 5/7-shot prompting therefore exhibits the optimal balance between translation quality and low error rates. In the news domain, GPT-4o 7-shot yields the highest overall system score and the lowest total error rate, though gains over an already high zero-shot baseline remain small. Across families, GPT 7-shot dominates in news with a system score of 0.915 and the lowest total and critical error rates by a wide margin, and GPT 5/7/10-shot configurations consistently outperform all LLaMa and NMT variants in the literary domain. However, fine-tuning catastrophically degrades GPT performance, with severe drops in system scores and spikes in error rates, a finding consistent with the other translation quality metrics, underscoring the risks of fine-tuning high-capacity LLMs with small datasets.

Key finding: LLaMa improves under adaptation but lags GPT

LLaMa benefits from both few-shot prompting and fine-tuning, with gradual reductions in major and total error rates, yet remains a middle-tier family behind GPT.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

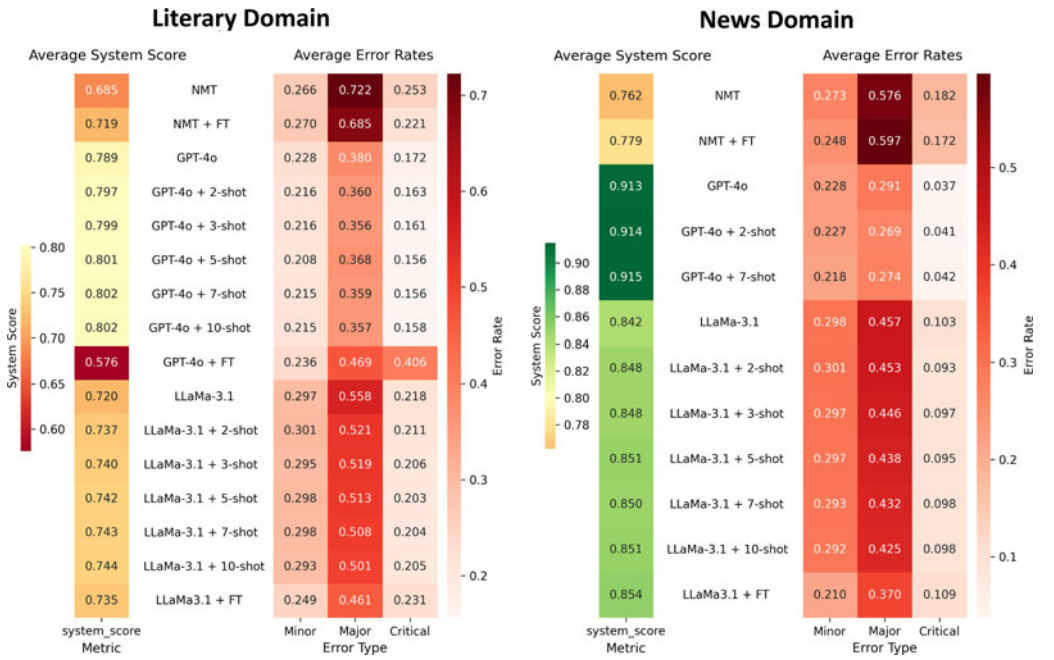


Figure 9. XCOMET system scores and error rate heatmaps for the literary and news domains. Only errors with higher than 50% confidence are reported.

LLaMa-3.1 shows gradual gains in the literary domain with increasing shot counts, accompanied by steady reductions in major and total error rates, and its fine-tuned variant further lowers major error rates while lifting the XCOMET system score with respect to pretrained LLaMa. In the news domain also, LLaMa gradually benefits from increasing shot counts (strongest at 10-shot) and reaches its best overall configuration under fine-tuning. Despite these gains, LLaMa still trails GPT in both system score and error rates, where fine-tuned LLaMa narrows but does not close the gap.

Key finding: NMT remains the weakest family despite fine-tuning

Fine-tuning delivers modest gains for NMT in both domains, but its system scores and error rates remain consistently below those of LLaMa and GPT.

NMT fine-tuning produces measurable reductions in total error rates and improves the XCOMET system score, most visibly in the literary domain. However, the improvements are limited in magnitude, leaving NMT as the weakest family in both domains with the lowest scores and highest error rates even after fine-tuning. The overall performance hierarchy from XCOMET therefore follows GPT $\geq$ LLaMa $\geq$ NMT, confirming the ranking observed in the translation quality metrics.

5.2.6 Correlation analysis

In this section, we present a detailed correlation analysis about the evaluation metrics we used in this work to better understand the relationships between the metrics. We compute Spearman rank correlations both within the three evaluation metric categories (translation quality, lexical diversity, and translation literalness) and across categories. This allows us to examine not only how metrics of the same type align with one another but also how different dimensions of translation

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

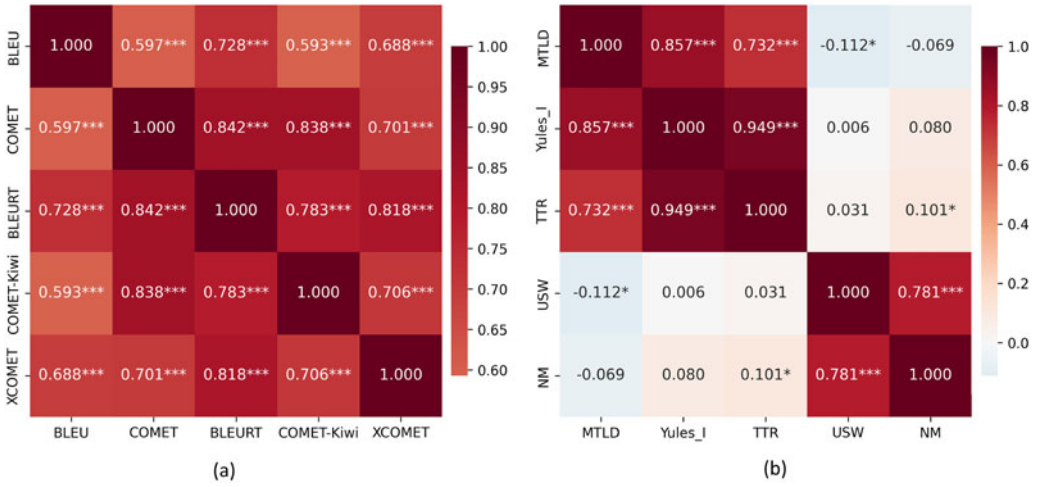


Figure 10. Spearman correlations (a) within translation quality metrics and (b) within lexical diversity and translation literalness metrics. Statistical significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

evaluation interact. In this way, we can observe the trade-offs between accuracy and creativeness of translations. The correlation computation involves all model configuration results across all language pairs. This amounts to 14 language directions (see Figure 1) and 16 model configurations (see Table 3) in the literary domain (224 in total) and 18 language directions (see Figure 2) and 12 model configurations (see Table 4) in the news domain (216 in total). We further report the confidence of these correlations by reporting statistical significance levels for all correlation coefficients. Within category correlation results are presented by concatenating the results of the two domains (440 translation evaluations) since both domains show similar trends. Between category correlation results are presented separately for each domain since we observe different trends for the two domains.

Key finding: BLEURT bridges surface and semantic quality metrics

All translation quality metrics correlate positively, but BLEURT occupies a central position, correlating strongly with surface-based and neural metrics.

Figure 10(a) shows the correlation matrix between the translation quality metrics. While all metrics share statistically significant positive associations, their strengths vary, reflecting both overlap and divergence in how they capture translation quality. BLEURT correlates strongly with both BLEU (0.728) and the neural metrics (COMET 0.842, XCOMET 0.818, COMET-Kiwi 0.783), bridging form- and meaning-oriented evaluation. COMET and COMET-Kiwi share a tight tie (0.838), indicating the shared underlying model assumptions. BLEU's alignment with learned metrics is weakest for COMET-Kiwi (0.593), indicating that surface matching is less compatible with reference-free semantic scoring. XCOMET aligns most with BLEURT (0.818), but has weaker ties to COMET (0.701) and BLEU (0.688).

Key finding: Diversity and literalness are independent dimensions

Lexical diversity and literalness metrics correlate strongly within their respective groups, but the two groups remain largely uncorrelated—capturing distinct aspects of translation behavior.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

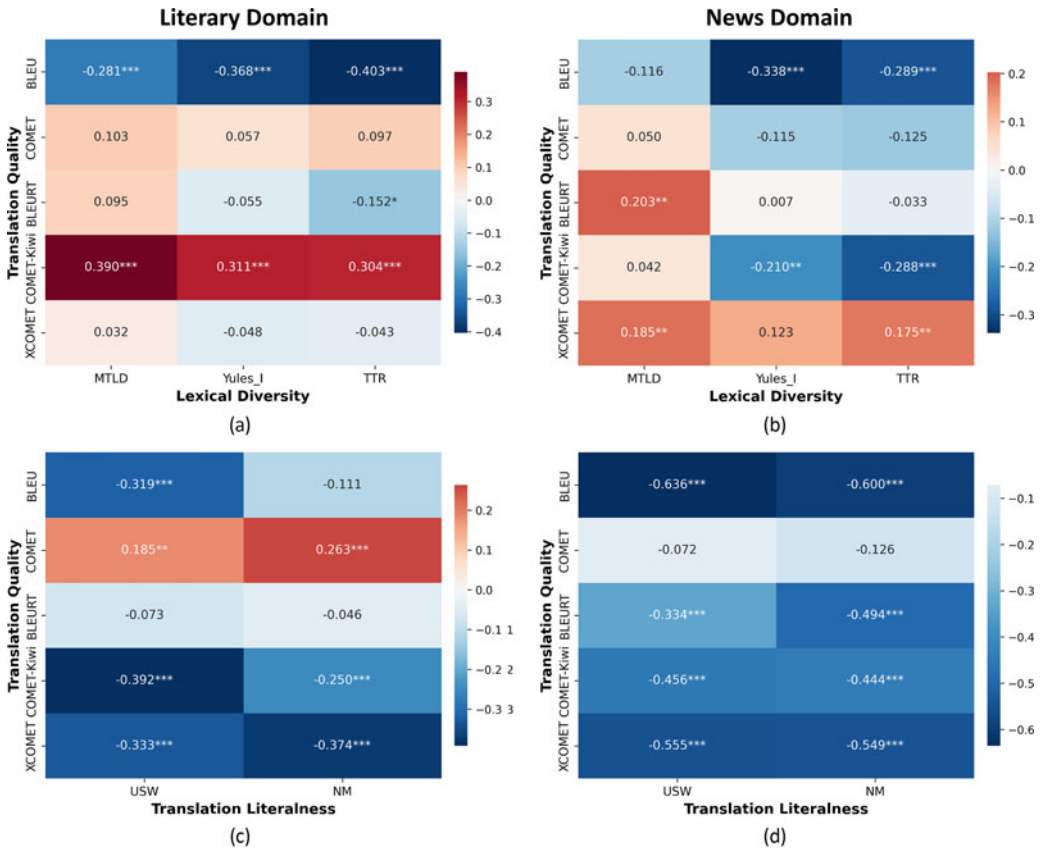


Figure 11. (Above) Spearman correlations between translation quality metrics and lexical diversity metrics (a) for the literary domain and (b) for the news domain. (Below) Spearman correlations between translation quality metrics and translation literalness metrics (c) for the literary domain and (d) for the news domain. Statistical significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Figure 10(b) shows that within the lexical diversity group, Yule's I and TTR are almost perfectly correlated (0.949) and both align strongly with MTLD (0.857 and 0.732). Within literalness, USW and NM share a strong positive correlation (0.781). Across the two groups, however, correlations are weak or slightly negative, confirming that lexical variety and structural/semantic freedom from the source are largely independent dimensions of translation behavior.

Key finding: BLEU penalizes diversity; semantic metrics reward it especially in literary

BLEU correlates negatively with lexical diversity in both domains, while semantic quality metrics favor richer vocabulary—more strongly in the literary than in the news domain. MTLD shows more consistent positive links to semantic quality.

When we look at the correlations between quality and diversity metrics in the literary domain (Figure 11a), we see that BLEU correlates negatively with all three diversity metrics, suggesting that more diverse translations tend to have less n-gram overlap with the reference translation. COMET-Kiwi correlates significantly positively with diversity metrics, indicating nonreference-based evaluation rewards translations with richer vocabulary and more varied lexical choices in the literary domain. In the news domain (Figure 11b), BLEU again penalizes

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table 7. Ablation Study: Impact of Training Budget on Translation Quality (Literary Domain). Local GPU costs calculated at 0.31 USD/hour. GPT cost based on $25/1M training tokens

Model	Training setting	BLEU	COMET	Est. cost (USD)
NMT (OPUS)	Baseline (Pre-trained)	9.65	0.766	–
	Fine-tuned (500k tokens)	9.92	0.768	<$0.01
	Fine-tuned (Full)	**14.21**	**0.787**	∼ $1.00
LLaMa-3.1	Zero-shot	5.97	0.735	–
	Fine-tuned (500k tokens)	6.90	0.742	<$0.02
	Fine-tuned (Full)	**13.43**	**0.783**	∼ $2.00
GPT-4o	Zero-shot	**9.51**	**0.786**	–
	Fine-tuned (500k tokens)	5.99	0.721	∼ $12.50

diversity (Yule's I −0.338, TTR −0.289), but semantic quality gains are weaker and COMET-Kiwi shows neutral or negative correlation. Across both domains, MTLD exhibits stronger positive links to semantic quality metrics than Yule's I and TTR, reflecting its robustness to text length.

Key finding: Domain-dependent trend for literalness

In the literary domain, COMET rewards freer (less literal) translations while COMET-Kiwi and XCOMET favor literalness. In the news domain, almost all quality metrics (BLEU, BLEURT, COMET-Kiwi, XCOMET) align with literalness, with COMET as the main outlier.

In the literary domain (Figure 11c), COMET correlates positively with both USW (0.185) and NM (0.263), favoring less literal translations, while COMET-Kiwi and XCOMET correlate significantly negatively with both metrics (−0.250 to −0.392), favoring source-aligned outputs; BLEURT is essentially neutral. In the news domain (Figure 11d) we observe a sharper pattern: BLEU, BLEURT, COMET-Kiwi, and XCOMET show strong negative correlations with nonliteralness, while COMET remains an outlier with weakly negative and nonsignificant correlations. Taken together, the results show a domain-dependent trend. In the literary domain, COMET favors freer translations while COMET-Kiwi and XCOMET prefer literal translations. In the news domain, most metrics (BLEU, BLEURT, COMET-Kiwi, XCOMET) favor more literal, word-by-word translations, with COMET as the main outlier.

5.2.7 Ablation Study: Impact of Training Dataset Size

As detailed in Section 5.1.1, GPT-4o fine-tuning was restricted to a 500k-token subset (≈5,000 samples). To evaluate the impact of this asymmetric training budget, we conducted an equalized-budget experiment on the English–Turkish language pair in the literary domain. We fine-tuned the open-source systems (NMT and LLaMa-3.1) on the exact 500k-token subset used for GPT-4o. The results, summarized in Table 7 and Figure 12, reveal a significant divergence in how these model architectures respond to small-scale adaptation.

The data show that while NMT and LLaMa-3.1 exhibit marginal gains (+0.27 and + 0.93 BLEU, and + 0.002 and + 0.007 COMET, respectively) when trained on 500k tokens, GPT-4o suffers from a severe performance degradation, with BLEU scores falling from 9.51 to 5.99 and COMET scores from 0.786 to 0.721. This decline occurred despite the fact that our training volume (5,000 samples) significantly exceeds the best practices suggested by the official OpenAI fine-tuning

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

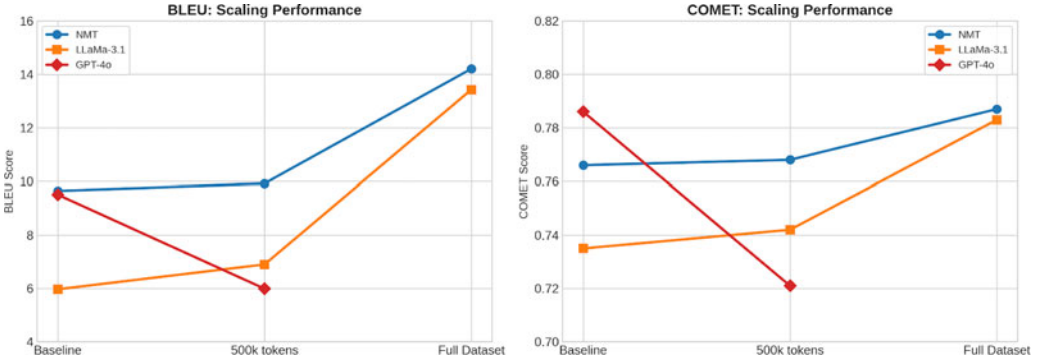


Figure 12. Ablation Study: Impact of Training Budget on Translation Quality (Literary Domain).

guidelines, which state that improvements are often observed with as few as 50–100 examples (OpenAI 2025).

This suggests that frontier models like GPT-4o are highly sensitive to weight updates on specialized datasets. The observed degradation is likely a result of catastrophic forgetting, where the model's general translation fluency is compromised by the specific stylistic patterns of the small-scale literary subset. Furthermore, GPT-4o fine-tuning proved to be the least cost-effective method: at $12.50 per 500k tokens, it was over 600 times more expensive than the local fine-tuning of LLaMa or NMT, yet yielded the lowest quality.

These findings suggest that simply increasing the GPT-4o fine-tuning budget probably would not necessarily recover performance in this setting. Frontier models appear to be optimized for in-context learning (few-shot prompting), whereas smaller open-source models demonstrate superior scalability through traditional fine-tuning—achieving substantial gains (+7.46 BLEU for LLaMa) when the full dataset is utilized. Consequently, we chose to prioritize our computational resources on these more responsive and cost-effective architectures.

6. Findings

In this section, we present our most key findings about the three dimensions we used to assess the quality of MT models. We also summarize our findings about the error analysis and the correlation analysis between the metrics.

Translation quality

- In the literary domain, **fine-tuned LLaMa** performs best for the BLEU metric, while **GPT** leads for the other metrics; in the news domain, **GPT zero- and few-shot** dominate in all translation quality metrics.
- **NMT fine-tuning** yields strong and consistent gains over pretrained NMT in the literary domain, but much weaker and inconsistent improvements in news.
- **LLaMa** benefits from both few-shot prompting and fine-tuning, with fine-tuning producing its largest translation quality improvements, especially for EN→X.
- **GPT few-shot** is highly effective in the literary domain but largely redundant in news; fine-tuning consistently harms GPT performance. In-context learning for GPT is more effective and cheaper than fine-tuning.
- **Across metrics**, BLEU favors fine-tuned NMT and LLaMa in the literary domain, while COMET, BLEURT, and COMET-Kiwi consistently favor GPT across domains.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

- **Among open source models**, fine-tuning or few-shot prompting the LLaMa model yields better translation quality than fine-tuning an NMT model (13/14 language directions for COMET in the literary domain, 16/18 in the news domain).
- **Human translations** achieve much higher BLEU, COMET, and BLEURT scores when compared to any model setting. Unlike other metrics, the **nonreference-based COMET-Kiwi** often rates MT systems above human translations—especially in the literary domain—highlighting its tendency to reward fluency and coherence over stylistic variation, with GPT models consistently outperforming all others.
- **Overall**, the performance hierarchy is **GPT** $\geq$ **LLaMa** $\geq$ **NMT**, with few-shot prompting emerging as the most efficient and reliable strategy for GPT; and fine-tuning for LLaMa, especially for non-English target languages.

Lexical diversity

- **GPT** with few-shot prompting produces the most lexically diverse translations in both domains, with significant gains over zero-shot in the news domain.
- **LLaMa** fine-tuning tends to reduce lexical diversity, suggesting a trade-off between translation quality and lexical richness for this model, while few-shot prompting yields limited and inconsistent benefits.
- **GPT** offers the strongest balance: high translation quality and high lexical diversity, especially under few-shot prompting.
- **GPT** few-shot **can surpass human** average MTLD in X→EN and, in the news domain, also in EN→X.

Translation literalness

- The **zero-shot** setting for LLMs produces relatively literal translations.
- **Few-shot prompting** and **fine-tuning** shift models toward less literal translations, with the effect strongest for LLaMa, especially for non-English target languages.
- **Fine-tuned LLaMa** produces the **least literal translations**. This model can be more preferable for creative, less direct, and high-quality translations.
- Unlike lexical diversity, GPT-4o versions fall behind fine-tuned LLaMa, which produces freer, less literal outputs.
- **USW and NM capture complementary aspects** of literalness, but both confirm the same trend: adaptation methods (few-shot, fine-tuning) reduce literalness for most language pairs in both domains.
- **Humans** generally produce less literal translations than MT systems across domains, though in specific directions fine-tuned LLaMa or NMT models yield less literal outputs, with LLaMa leading for X→EN in the news domain.

Error analysis

- In both domains, GPT with few-shot prompting achieves the best balance of translation quality and low error rates.
- LLaMa benefits from few-shot prompting and fine-tuning, with gradual reductions in major and total error rates, though it still lags behind GPT.
- NMT gains modest improvements from fine-tuning in both domains but consistently underperforms compared to LLaMa and GPT.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Metric correlations

- Translation quality metrics largely agree, with **BLEURT as the central bridge** between surface-based (BLEU) and semantic (COMET family) measures.
- Lexical diversity metrics align closely with each other and literalness metrics correlate significantly with each other, but the two groups remain largely independent, signaling they capture different aspects of translation.
- **BLEU consistently penalizes lexical diversity**, and in the news domain also strongly **favors literal translations**.
- Semantic metrics (BLEURT, COMET, COMET-Kiwi, XCOMET) diverge by domain:
  - In the **literary domain**, COMET rewards freer, less literal translations, while COMET-Kiwi and XCOMET prefer literalness.
  - In the **news domain**, COMET remains relatively neutral, but BLEURT, COMET-Kiwi, and XCOMET strongly align with literalness.
- Among lexical diversity metrics, MTLD shows stronger positive links to semantic quality metrics.
- **Domain effect:** diversity is rewarded in literary translation, while literalness is rewarded in news, underscoring domain-dependent trade-offs between accuracy and creativity.

7. Conclusion

In this work, we conducted a comparative analysis of traditional NMT models and LLMs from three dimensions: translation quality, lexical diversity, and translation literalness. We identified distinct strengths and challenges associated with each model. The commercial GPT-4o model particularly stood out, demonstrating superior translation quality and lexical diversity, especially with in-context learning methods like few-shot prompting. This model consistently achieved high COMET, BLEURT, and COMET-Kiwi scores across most language pairs, and our error analysis confirmed it provides the best balance of high quality and low error rates. Remarkably, few-shot GPT-4o was even able to surpass human average lexical diversity in specific translation directions, illustrating its proficiency in producing contextually and lexically rich translations without additional fine-tuning.

We also uncovered surprising limitations of fine-tuning depending on the model architecture. While fine-tuning improved the translation performance of open-source models such as NMT and LLaMa, it consistently harmed the performance of the commercial GPT-4o model. Furthermore, fine-tuning led to a notable decline in lexical diversity for LLaMa, revealing a trade-off between adaptation and lexical richness. On the other hand, fine-tuning successfully resulted in more liberal outputs, especially with the LLaMa model delivering high-quality translations that are less literal, suitable for contexts demanding creative interpretation.

Our analysis of translation evaluation metrics further revealed systematic differences. BLEU favored fine-tuned NMT and LLaMa in the literary domain, while COMET consistently favored GPT across domains, underscoring the distinction between surface-oriented and semantic-oriented evaluation. Interestingly, when compared to human baselines, the nonreference-based COMET-Kiwi metric occasionally rated MT systems—especially GPT—above human translations, highlighting its tendency to reward fluency over stylistic variation. Correlation analysis showed that lexical diversity and translation literalness are largely independent dimensions: BLEU penalizes lexical diversity and favors literalness, while semantic metrics diverge by domain. In literary texts, COMET rewarded freer, less literal translations, whereas COMET-Kiwi and XCOMET preferred literalness. In the news domain, nearly all the metrics favor literalness.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Taken together, our findings suggest that selecting between different MT systems should be driven by specific translation objectives. For lexically rich and high-quality translations, few-shot GPT-4o is most effective; for less literal, more creative translations, fine-tuned LLaMa is more suitable. We also highlight a clear domain effect: literary translation benefits from lexical diversity, while news translation rewards literalness, reflecting different trade-offs between creativity and accuracy.

These insights pave the way for future research to explore hybrid approaches, combining the strengths of these models to tackle complex linguistic tasks effectively. Future work may also include human evaluation on how translation quality correlates with literalness and creativity, in addition to exploring different decoding or policy optimization methods to control lexical richness and literalness. Furthermore, extending this framework to low-resource language pairs will be a critical next step to investigate whether the observed trends in domain adaptation and architectural sensitivity generalize beyond high-resource settings.

Acknowledgements. The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Funding statement. This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) (grant number 121K221) and the Boğaziçi University Research Fund (grant number 19986).

Competing interests. The authors declare none.

Data availability statement. Dataset splits (except for Turkish for copyright issues), the code, and all evaluation results are available on our Github page.[n]

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

AI usage. We use the Cursor[o] tool with *claude-4-sonnet*[p] to help us write the evaluation and data processing Python scripts. All code has been carefully reviewed. ChatGPT[q] was utilized to spell-check and correct the grammar of this document. The resulting manuscript has been carefully reviewed by the authors for correctness, flow, and articulation.

References

Alves D., Guerreiro N., Alves J., Pombal J., Rei R., de Souza J., Colombo P. and Martins A. (2023). Steering large language models for machine translation with finetuning and in-context learning. In Bouamor H., Pino J. and Bali K. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 11127–11148.

Baziotis C., Mathur P. and Hasler E. (2023). Automatic evaluation and analysis of idioms in neural machine translation. In Vlachos A. and Augenstein I. (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3682–3700.

Bentivogli L., Bisazza A., Cettolo M. and Federico M. (2016). Neural versus phrase-based machine translation quality: A case study. In Su J., Duh K. and Carreras X. (eds), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 257–267.

Dankers V., Lucas C. and Titov I. (2022). Can transformer be too compositional? Analysing idiom processing in neural machine translation. In Muresan S., Nakov P. and Villavicencio A. (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3608–3626.

Dettmers T., Pagnoni A., Holtzman A. and Zettlemoyer L. (2023). QLoRA: Efficient finetuning of quantized LLMs.

[n] https://github.com/zeynepyirmibes/diversity_literalness_mt
[o] https://cursor.com/
[p] https://www.anthropic.com/claude/sonnet
[q] https://chatgpt.com/

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Freitag M. and Al-Onaizan Y. (2017). Beam search strategies for neural machine translation. In Luong T., Birch A., Neubig G. and Finch A. (eds), *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, pp. 56–60.

Freitag M., Vilar D., Grangier D., Cherry C. and Foster G. (2022). A natural diet: Towards improving naturalness of machine translation output. In Muresan S., Nakov P. and Villavicencio A. (eds), *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 3340–3353.

Frontull S. and Ströhle T. (2025). Compensating for data with reasoning: Low-resource machine translation with LLMs.

Garcia X., Bansal Y., Cherry C., Foster G., Krikun M., Johnson M. and Firat O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.

Guerreiro N.M., Rei R., Stigt D., Coheur L., Colombo P. and Martins A.F.T. (2024). xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics* **12**, 979–995.

Hendy A., Abdelrehim M., Sharaf A., Raunak V., Gabr M., Matsushita H., Kim Y.J., Afify M. and Awadalla H.H. (2023). How good are GPT models at machine translation? A comprehensive evaluation.

Holtzman A., Buys J., Du L., Forbes M. and Choi Y. (2020). The curious case of neural text degeneration.

Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L. and Chen W. (2021). LoRA : Low-rank adaptation of large language models.

Huang Y., Li B., Feng X., Huo W., Fu C., Liu T. and Qin B. (2024). Aligning translation-specific understanding to general understanding in large language models. In Al-Onaizan Y., Bansal M. and Chen Y.-N. (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 5028–5041.

Jiao W., Wang W., tse Huang J., Wang X., Shi S. and Tu Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine.

Kocmi T., Avramidis E., Bawden R., Bojar O., Dvorkovich A., Federmann C., Fishel M., Freitag M., Gowda T., Grundkiewicz R., Haddow B., Karpinska M., Koehn P., Marie B., Monz C., Murray K., Nagata M., Popel M., Popović M., Shmatova M., Steingrímsson S. and Zouhar V. (2024a). Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Haddow B., Kocmi T., Koehn P. and Monz C. (eds), *Proceedings of the Ninth Conference on Machine Translation*. Miami, Florida, USA: Association for Computational Linguistics, pp. 1–46.

Kocmi T., Avramidis E., Bawden R., Bojar O., Dvorkovich A., Federmann C., Fishel M., Freitag M., Gowda T., Grundkiewicz R., Haddow B., Koehn P., Marie B., Monz C., Morishita M., Murray K., Nagata M., Nakazawa T., Popel M., Popović M. and Shmatova M. (2023). Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn P., Haddow B., Kocmi T. and Monz C. (eds), *Proceedings of the Eighth Conference on Machine Translation*. Singapore: Association for Computational Linguistics, pp. 1–42.

Kocmi T., Zouhar V., Federmann C. and Post M. (2024b). Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Ku L.-W., Martins A. and Srikumar V. (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 1999–2014.

Martínez G., Hernández J.A., Conde J., Reviriego P. and **Merino-Gómez E.** (2024). Beware of words: Evaluating the lexical diversity of conversational LLMs using chatGPT as case study. *ACM Transactions on Intelligent Systems and Technology* **16**(6), 138.

McCarthy P.M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). PhD thesis, Copyright - Database copyright ProQuest LLC. ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-08-04.

Moslem Y., Haque R., Kelleher J.D. and Way A. (2023). Adaptive machine translation with large language models. In Nurminen M., Brenner J., Koponen M., Latomaa S., Mikhailov M., Schierl F., Ranasinghe T., Vanmassenhove E., Vidal S.A., Aranberri N., Nunziatini M., Escartín C.P., Forcada M., Popovic M., Scarton C. and Moniz H. (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Tampere, Finland: European Association for Machine Translation, pp. 227–237.

OpenAI. (2025). Supervised fine-tuning. OpenAI Platform Documentation. Retrieved from https://developers.openai.com/api/docs/guides/supervised-fine-tuning [Accessed October 2025].

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle P., Charniak E. and Lin D. (eds), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.

Peng K., Ding L., Zhong Q., Shen L., Liu X., Zhang M., Ouyang Y. and Tao D. (2023). Towards making the most of ChatGPT for machine translation. In Bouamor H., Pino J. and Bali K. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 5622–5633.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Ploeger E., **Lai H.**, **Van Noord R. and Toral A.** (2024). Towards tailored recovery of lexical diversity in literary machine translation. In Scarton C., Prescott C., Bayliss C., Oakley C., Wright J., Wrigley S., Song X., Gow-Smith E., Bawden R., Sánchez-Cartagena V.M., Cadwell P., Lapshinova-Koltunski E., Cabarrão V., Chatzitheodorou K., Nurminen M., Kanojia, D. and Moniz H. (eds), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. Sheffield, UK: European Association for Machine Translation (EAMT), pp. 286–299.

Post M. (2018). A call for clarity in reporting BLEU scores. In Bojar O., Chatterjee R., Federmann C., Fishel M., Graham Y., Haddow B., Huck M., Yepes A.J., Koehn P., Monz C., Negri M., Névéol A., Neves M., Post M., Specia L., Turchi M. and Verspoor K. (eds), *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 186–191.

Raunak V., **Menezes A.**, **Post M. and Hassan H.** (2023). Do GPTs produce less literal translations? In Rogers A., Boyd-Graber J. and Okazaki N. (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 1041–1050.

Rei R., **Stewart C.**, **Farinha A.C. and Lavie A.** (2020). COMET: A neural framework for MT evaluation. In Webber B., Cohn T., He Y. and Liu Y. (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 2685–2702.

Rei R., **Treviso M.**, **Guerreiro N.M.**, **Zerva C.**, **Farinha A.C.**, **Maroti C.**, **de Souza J.G.C.**, **Glushkova T.**, **Alves D.M.**, **Lavie A.**, **Coheur L. and Martins A.F.T.** (2022). CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task.

Reviriego P., **Conde J.**, **Merino-Gómez E.**, **Martínez G. and Hernández J.A.** (2024). Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications* **18**, 100602.

Schioppa A., **Vilar D.**, **Sokolov A. and Filippova K.** (2021). Controlling machine translation for multiple attributes with additive interventions. In Moens M.-F., Huang X., Specia L. and Yih S.W. (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6676–6696.

Sellam T., **Das D. and Parikh A.** (2020). BLEURT: Learning robust metrics for text generation. In Jurafsky D., Chai J., Schluter N. and Tetreault J. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 7881–7892.

Shen L. (2022). LexicalRichness: A small module to compute textual lexical richness.

Song Y., **Li L.**, **Lothritz C.**, **Ezzini S.**, **Sleem L.**, **Gentile N.**, **State R.**, **Bissyandé T.F. and Klein J.** (2025). Is small language model the silver bullet to low-resource languages machine translation?

Stap D., **Hasler E.**, **Byrne B.**, **Monz C. and Tran K.** (2024). The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In Ku L.-W., Martins A. and Srikumar V. (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, pp. 6189–6206.

Su T., **Peng X.**, **Thillainathan S.**, **Guzmán D.**, **Ranathunga S. and Lee E.-S.** (2024). Unlocking parameter-efficient fine-tuning for low-resource language translation. In Duh K., Gomez H. and Bethard S. (eds), *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 4217–4225.

Tang L., **Qin J.**, **Ye W.**, **Tan H. and Yang Z.** (2025). Adaptive few-shot prompting for machine translation with pre-trained language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'25/IAAI'25/EAAI'25*. AAAI Press.

Templin M.C. (1957). *Certain Language Skills in Children: Their Development and Interrelationships*, vol. **26**. ned - new edition , Edn. University of Minnesota Press.

Tiedemann J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2214–2218.

Tiedemann J. (2020). The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, pp. 1174–1182. Online.

Touvron H., **Martin L.**, **Stone K.**, **Albert P.**, **Almahairi A.**, **Babaei Y.**, **Bashlykov N.**, **Batra S.**, **Bhargava P.**, **Bhosale S.**, **Bikel D.**, **Blecher L.**, **Ferrer C.C.**, **Chen M.**, **Cucurull G.**, **Esiobu D.**, **Fernandes J.**, **Fu J.**, **Fu W.**, **Fuller B.**, **Gao C.**, **Goswami V.**, **Goyal N.**, **Hartshorn A.**, **Hosseini S.**, **Hou R.**, **Inan H.**, **Kardas M.**, **Kerkez V.**, **Khabsa M.**, **Kloumann I.**, **Korenev A.**, **Koura P.S.**, **Lachaux M.-A.**, **Lavril T.**, **Lee J.**, **Liskovich D.**, **Lu Y.**, **Mao Y.**, **Martinet X.**, **Mihaylov T.**, **Mishra P.**, **Molybog I.**, **Nie Y.**, **Poulton A.**, **Reizenstein J.**, **Rungta R.**, **Saladi K.**, **Schelten A.**, **Silva R.**, **Smith E.M.**, **Subramanian R.**, **Tan X.E.**, **Tang B.**, **Taylor R.**, **Williams A.**, **Kuan J.X.**, **Xu P.**, **Yan Z.**, **Zarov I.**, **Zhang Y.**, **Fan A.**, **Kambadur M.**, **Narang S.**, **Rodriguez A.**, **Stojnic R.**, **Edunov S. and Scialom T.** (2023). Llama 2: Open foundation and fine-tuned chat models.

Vanmassenhove E., **Shterionov D. and Gwilliam M.** (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Merlo P., Tiedemann J. and Tsarfaty R. (eds), *Proceedings of the 16th Conference of the*

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

European Chapter of the Association for Computational Linguistics: Main Volume. Online. Association for Computational Linguistics, pp. 2203–2213.

Vanmassenhove E., **Shterionov D. and Way A.** (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In Forcada M., Way A., Haddow B. and Sennrich R. (eds), *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin, Ireland: European Association for Machine Translation, pp. 222–232.

Xu H., **Kim Y.J.**, **Sharaf A. and Awadalla H.H.** (2024). A paradigm shift in machine translation: Boosting translation performance of large language models.

Xu J., **Crego J. and Senellart J.** (2020). Boosting neural machine translation with similar translations. In Jurafsky D., Chai J., Schluter N. and Tetreault J. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1580–1590. Online.

Yang C., **Dou Y.**, **Heineman D.**, **Wu X. and Xu W.** (2025). Evaluating LLMs on Chinese idiom translation.

Yao G. and Fan L. (2025). An entropy-based study of simplification in ChatGPT translations compared to neural machine translation and human translation across genres. *PLoS ONE* **20**(12), e0339762.

Yirmibeşoğlu Z., **Dursun O.**, **Dalli H.**, **Şahin M.**, **Hodzik E.**, **Gürses S. and Güngör T.** (2023). Incorporating human translator style into English-Turkish literary machine translation. In Nurminen M., Brenner J., Koponen M., Latomaa S., Mikhailov M., Schierl F., Ranasinghe T., Vanmassenhove E., Vidal S.A., Aranberri N., Nunziatini M., Escartín C.P., Forcada M., Popovic M., Scarton C. and Moniz H. (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Tampere, Finland: European Association for Machine Translation, pp. 419–428.

Yule G. (1944). *The Statistical Study of Literary Vocabulary*. The University Press.

Zhang X., **Rajabi N.**, **Duh K. and Koehn P.** (2023). Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In Koehn P., Haddow B., Kocmi T. and Monz C. (eds), *Proceedings of the Eighth Conference on Machine Translation*. Singapore: Association for Computational Linguistics, pp. 468–481.

Appendix A. Lexical diversity metrics

We use the LexicalRichness module (Shen 2022) to compute the three lexical diversity metrics.

TTR is calculated as follows:

$$\mathrm{TTR} = \frac{V}{N} \tag{A.1}$$

Yule's I is calculated as follows:

$$I = \frac{V^2}{M_2 - V} \tag{A.2}$$

$$M_2 = \sum_{i=1}^{V} i^2 * f_v(i, N) \tag{A.3}$$

In these formulas, N is the total number of tokens in the text and V is the number of unique word types. $f_v(i, N)$ refers to the number of word types occurring i times in a sample of length N.

The pseudocode of MTLD is provided in Algorithm 1. The TTR threshold represents the point at which the text is considered to have a stable type-token ratio (TTR).

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Algorithm 1. Calculate MTLD

```
function mtld(text, threshold = 0.72)
    function calculate_ttr(text)
        return unique_tokens(text) / total_tokens(text)
    end function
    function calculate_mtld(text, min_words)
        initialize factor ← 0
        initialize factor_lengths ← 0
        initialize start ← 0
        for each word position x in text do
            initialize segment_text ← text[start : x + 1]
            if (x + 1 = length(text)) then
                factor ← factor + (1 - calculate_ttr(segment_text)) / (1 - threshold)
                factor_lengths ← factor_lengths + length(segment_text)
            else if (calculate_ttr(segment_text) <threshold
            and length(segment_text) ≥ min_words) then
                factor ← factor + 1
                factor_lengths ← factor_lengths + length(segment_text)
                start ← x + 1
        end for
        return factor_lengths / factor
    end function
    initialize reversed_text ← reverse(text)
    return (calculate_mtld(text, 10) + calculate_mtld(reversed_text, 10))/2
end function
```

Appendix B. Neural machine translation models and hyperparameters

Table B1. Pretrained neural machine translation models and their fine-tuning hyperparameters in the news domain. Model URLs follow the pattern: https://huggingface.co/Helsinki-NLP/{Model}

Direction	Model	Batch size	Epochs	Learning Rate
DE-EN	opus-mt-de-en	128	3	1×10^{-5}
EN-DE	opus-mt-en-de	128	3	1×10^{-5}
EN-ES	opus-mt-tc-big-en-es	128	3	1×10^{-5}
ES-EN	opus-mt-es-en	128	3	1×10^{-5}
EN-IT	opus-mt-tc-big-en-it	128	3	1×10^{-5}
IT-EN	opus-mt-tc-big-it-en	128	3	1×10^{-5}

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table B1. Continued

Direction	Model	Batch size	Epochs	Learning Rate
FR-EN	opus-mt-tc-big-fr-en	128	3	1×10^{-6}
EN-FR	opus-mt-tc-big-en-fr	128	3	2×10^{-5}
TR-EN	opus-mt-tc-big-tr-en	128	3	1×10^{-6}
EN-TR	opus-mt-tc-big-en-tr	128	3	1×10^{-6}
RU-EN	opus-mt-ru-en	128	3	1×10^{-5}
EN-RU	opus-mt-en-ru	128	3	2×10^{-5}
ZH-EN	opus-mt-zh-en	128	3	2×10^{-5}
EN-ZH	opus-mt-en-zh	128	3	2×10^{-5}
HI-EN	opus-mt-hi-en	128	3	1×10^{-5}
EN-HI	opus-mt-en-hi	128	3	2×10^{-5}
JA-EN	opus-mt-ja-en	128	3	2×10^{-5}
EN-JA	opus-tatoeba-en-ja	128	3	2×10^{-5}

Appendix C. WMT news test sets

Table C1. WMT news test sets for each language pair

Language pair	WMT test
DE-EN	newstest2021
ES-EN	newstest2013
FR-EN	newstest2014
HI-EN	newstest2014
IT-EN	newstest2009
JA-EN	newstest2021
RU-EN	newstest2021
TR-EN	newstest2018
ZH-EN	newstest2021

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Appendix D. BLEURT and COMET-Kiwi results

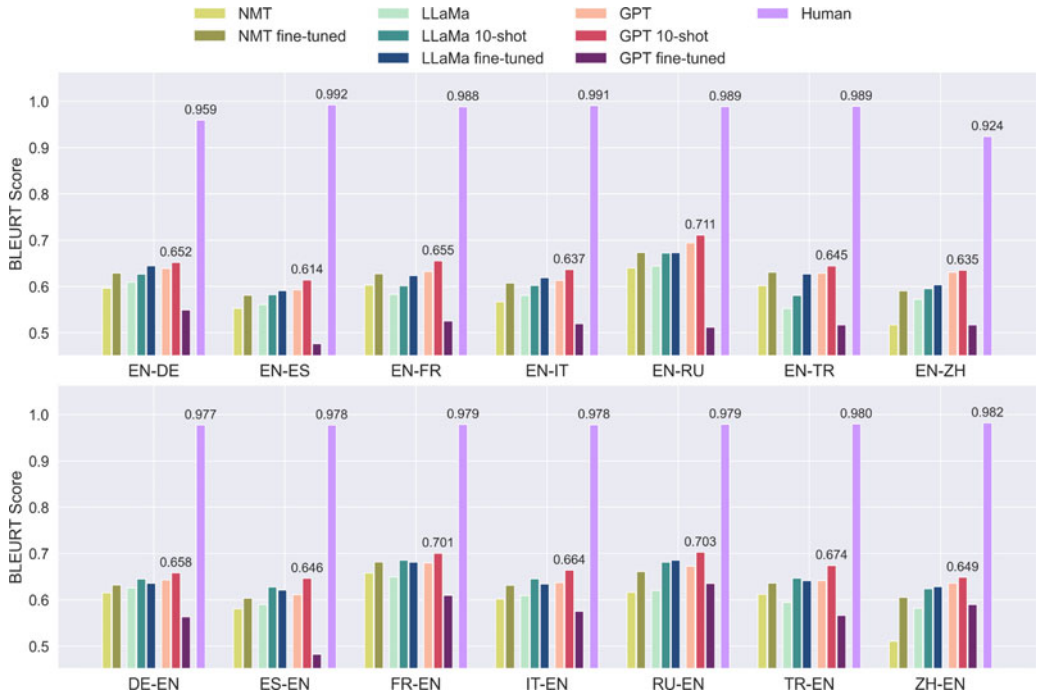


Figure D1. BLEURT scores of the models for English→X (above) and X→English (below) language pairs for the literary domain.

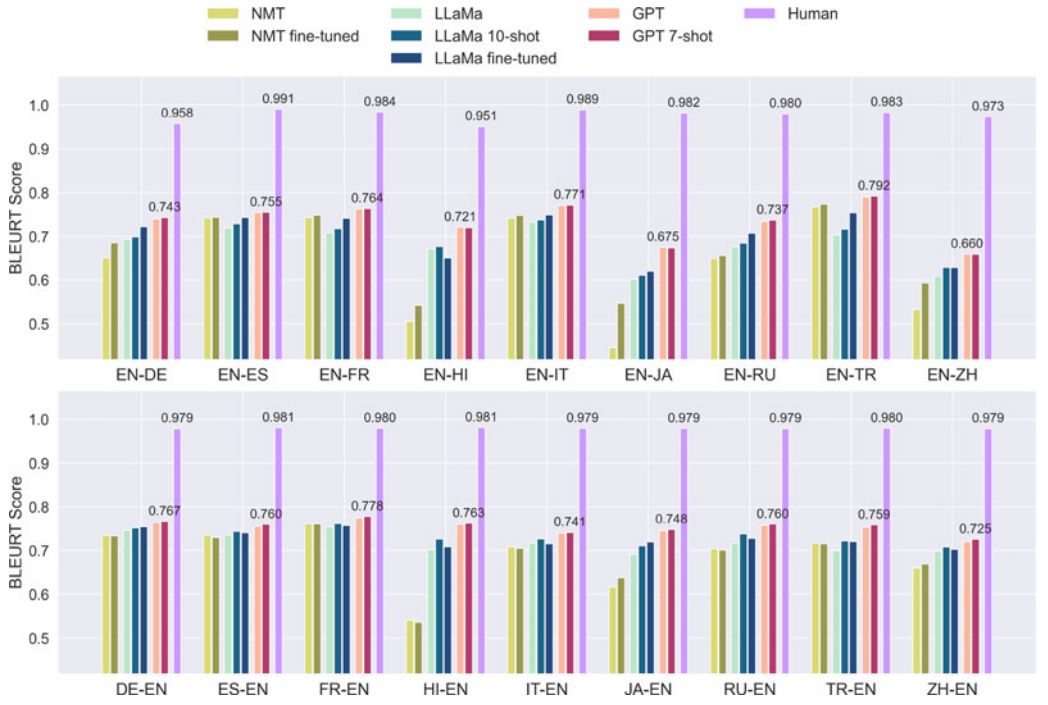


Figure D2. BLEURT scores of the models for English→X (above) and X→English (below) language pairs for the news domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

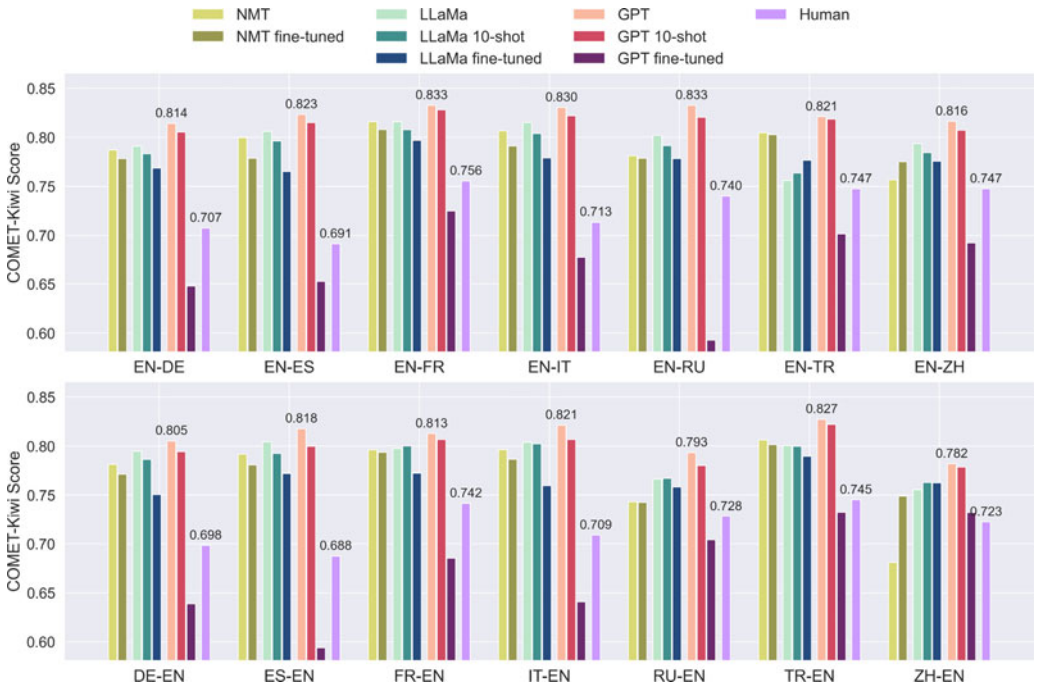


Figure D3. COMET-Kiwi scores of the models for English→X (above) and X→English (below) language pairs for the literary domain.

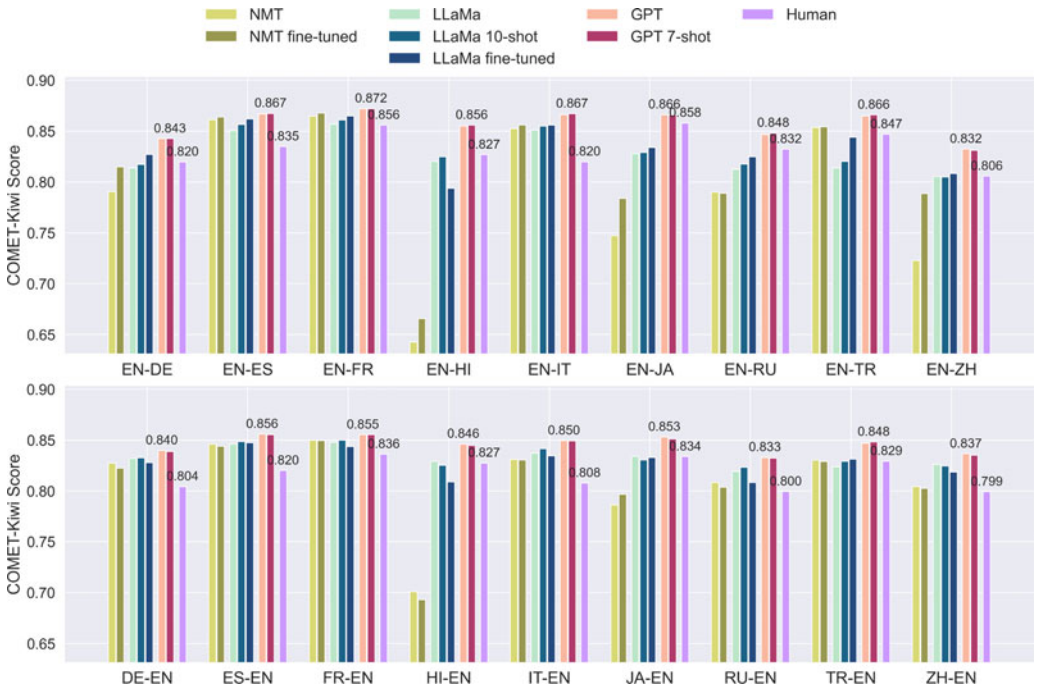


Figure D4. COMET-Kiwi scores of the models for English→X (above) and X→English (below) language pairs for the news domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Appendix E. TTR and Yule's I results

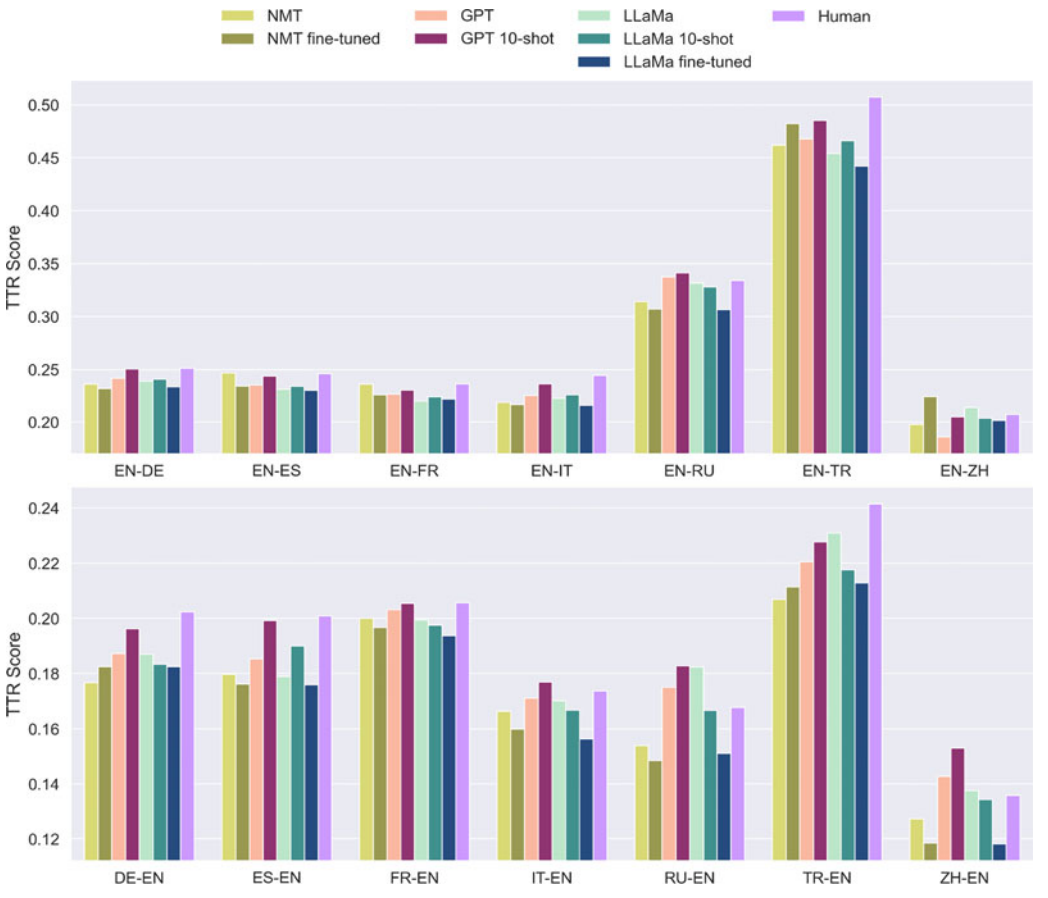


Figure E1. TTR scores of pretrained and best-performing models for the literary domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

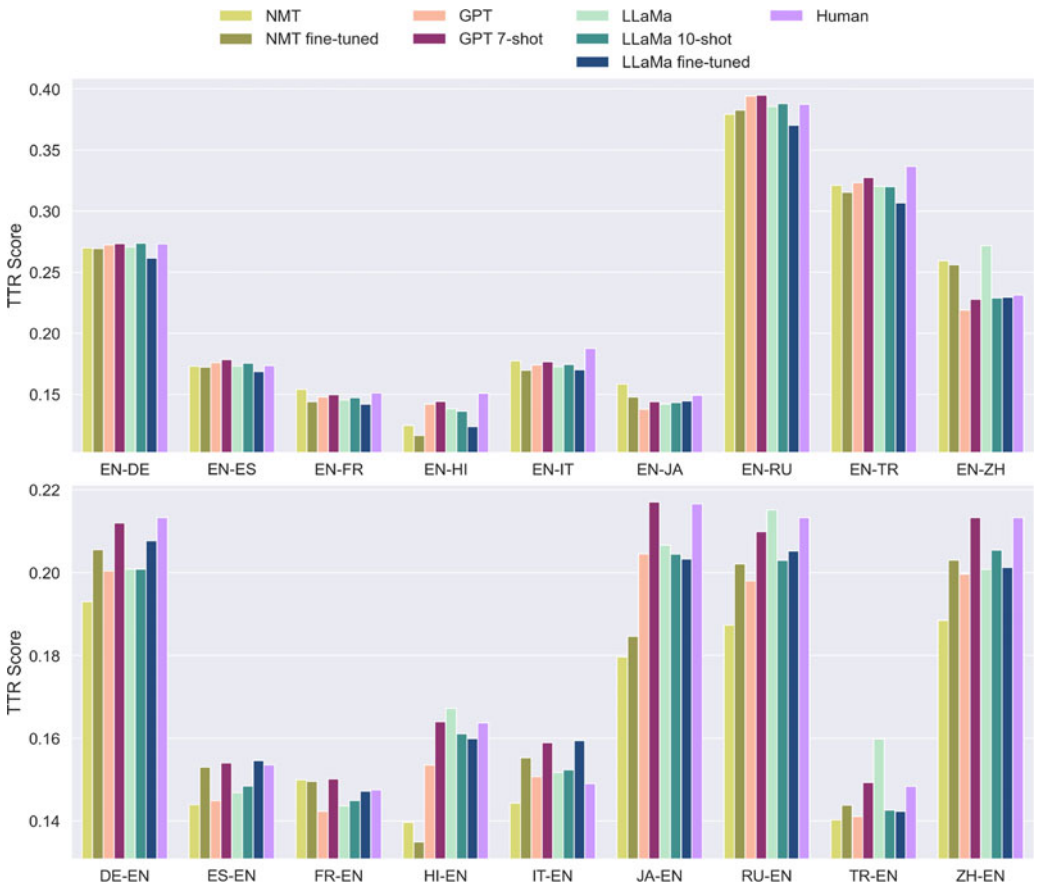


Figure E2. TTR scores of pretrained and best-performing models for the news domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

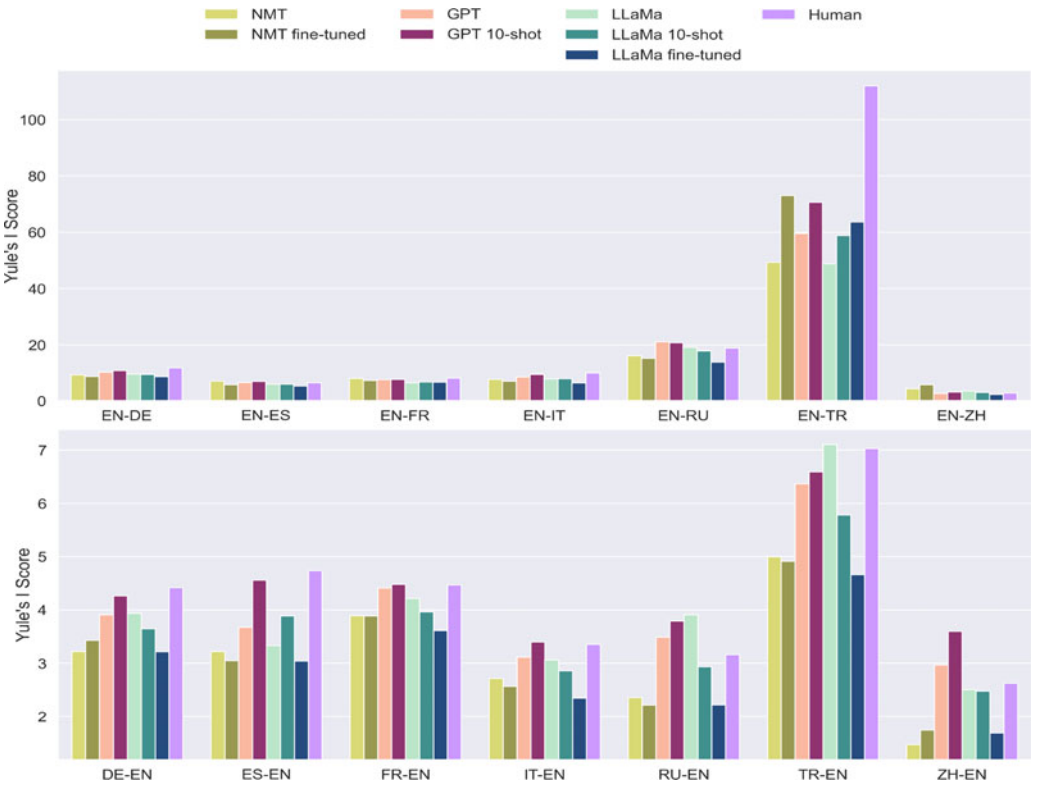


Figure E3. Yule's I scores of pretrained and best-performing models for the literary domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

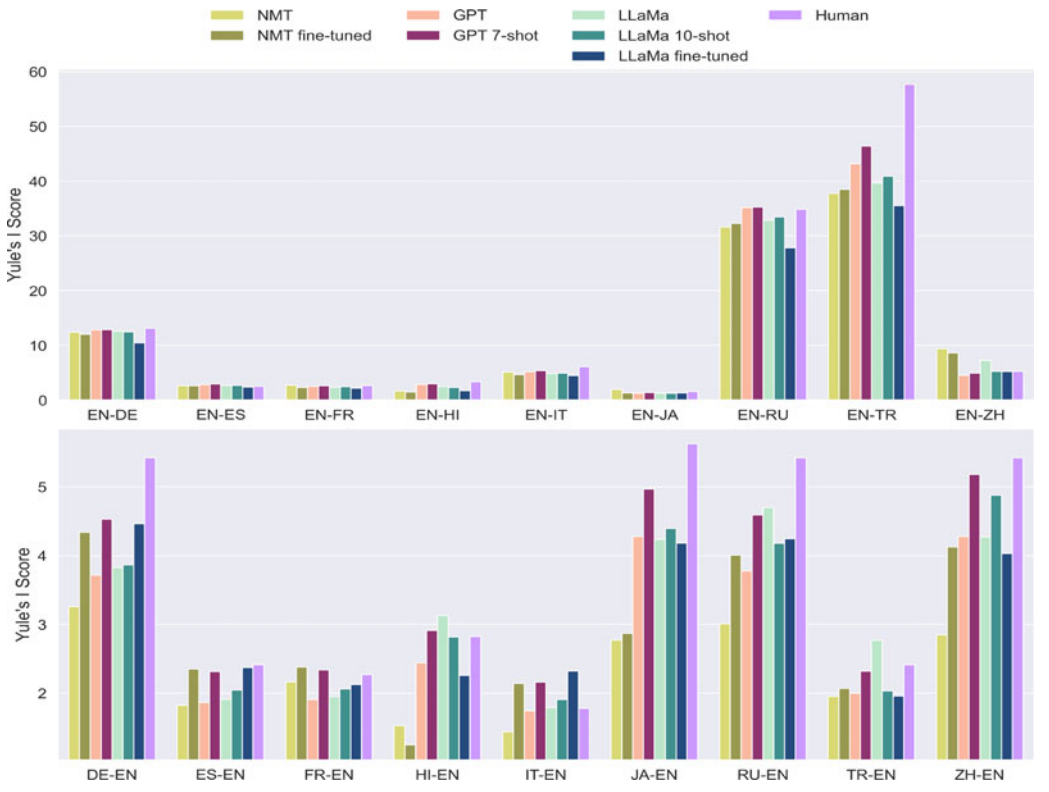


Figure E4. Yule's I scores of pretrained and best-performing models for the news domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Appendix F. USW and NM results

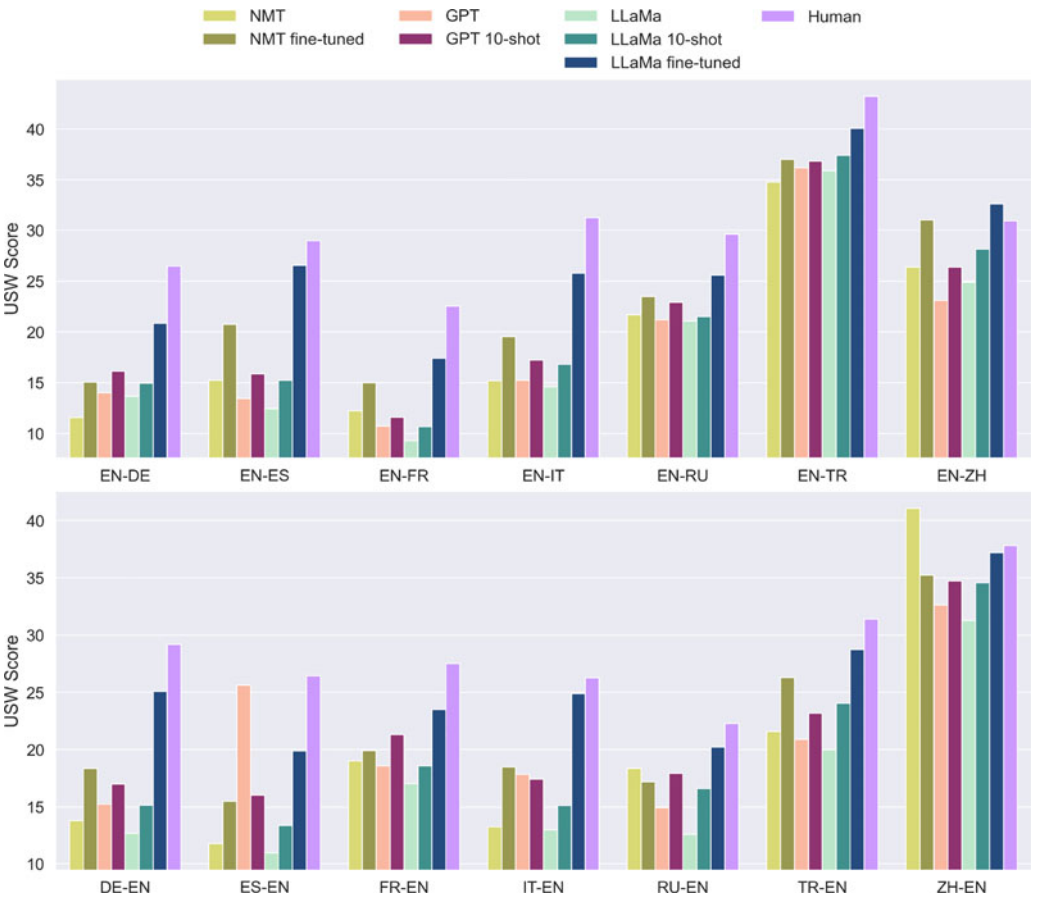


Figure F1. USW scores of pretrained and best-performing models for the literary domain.

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

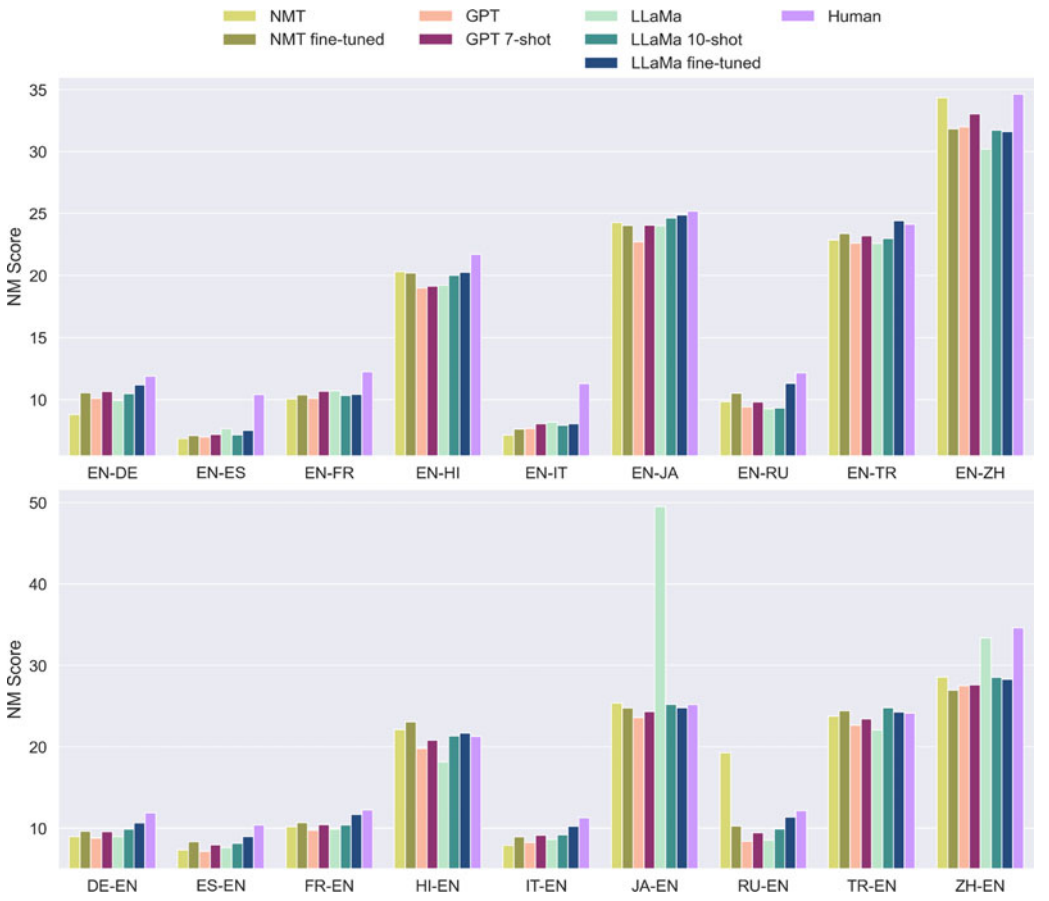


Figure F2. NM scores of pretrained and best-performing models for the news domain.

Appendix G. Directional asymmetries in translation Performance

Our evaluation of translation quality and lexical diversity reveals asymmetries between the EN→X and X→EN directions. Because these patterns may be misinterpreted as evidence of directional superiority, we dedicate this appendix to analyzing their likely causes.

Translation quality. The scores in Tables 3 and 4 exhibit a consistent direction-based asymmetry, with X→EN typically outperforming EN→X across several metrics. This gap is likely driven by a combination of factors. First, the predominant English-centric nature of most pretraining corpora (for LLMs) and parallel corpora (for NMT) tends to produce more robust language modeling in English, often resulting in higher fluency when English is the target language. Second, surface-based metrics like BLEU can disproportionately penalize target languages with higher linguistic complexity than English. This includes morphologically rich languages such as Turkish and Russian, where a lemma may appear in numerous inflected forms not present in the reference, as well as languages like Japanese where tokenization and segmentation nuances can lead to lower n-gram overlap. Notably, this asymmetry narrows or reverses under COMET and COMET-Kiwi in the literary domain, suggesting that the X→EN superiority is partly a property of surface-matching metrics rather than an underlying quality gap. Consequently, the apparently higher X→EN quality should be interpreted with caution, as it reflects both English-centric modeling advantages and reference-based metric biases, and is partly consistent with the possibility

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

that EN→X outputs are acceptable but diverge from a single reference. This in turn motivates the use of semantic metrics with different underlying assumptions (such as BLEURT together with COMET) and, where feasible, human evaluation to complement surface-based metrics.

Lexical diversity. The lexical diversity results in Tables 5 and 6 display the opposite asymmetry: EN→X translations achieve substantially higher TTR, Yule's I, and MTLD scores than X→EN translations for most configurations. Crucially, this asymmetry is not primarily driven by translation quality but is largely a linguistic artifact of the target languages. Morphologically rich and agglutinative languages such as Turkish produce numerous surface forms from a single lemma through suffixation, while languages like Russian exhibit similar inflation through complex inflectional paradigms. Furthermore, German significantly increases unique type counts through productive noun compounding, where complex concepts are fused into single tokens that would be expressed as multiple frequent words in English. These linguistic properties mechanically inflate type counts and, consequently, all type-based diversity measures. English, by contrast, has comparatively limited inflection and compounding, so X→EN translations naturally yield lower diversity scores regardless of model performance. The clearest evidence for this artifact is EN→TR in the literary domain, whose scores are so far above the rest that they required a separate plotting scale. It would therefore be inaccurate to infer from raw scores that EN→X is produced more creatively; equal-quality translations into morphologically complex or compounding languages will mechanically score higher on lexical diversity metrics.

Appendix H. Aggregate statistical significance of core findings

We evaluate statistical significance for a set of pairwise contrasts within each domain and translation direction, separating within-family comparisons from cross-family comparisons. These contrasts are specifically selected to evaluate our core research hypotheses and key findings. For each contrast, we computed paired differences between systems across five metrics: BLEU, COMET, MTLD, USW, and NM with metric-appropriate paired tests: sacreBLEU's paired bootstrap for BLEU, a paired segment-level t-test for COMET, paired bootstrap resampling over MTLD chunks, and paired permutation tests for sentence-level USW and NM. All tests are two-sided and report the score difference as system A minus system B, with 95% confidence intervals and raw p-values. To control for multiple comparisons, we applied Holm-Bonferroni correction separately within each domain, translation direction, metric, and contrast group, treating results with Holm-adjusted $p < 0.05$ as significant. Finally, to summarize patterns across translation directions, we counted significant wins, losses, and ties for each contrast and metric, and used a two-sided binomial sign test on the win/loss counts to assess whether the cross-system trend was stronger than expected by chance. Tables H1 and H2 summarize the pairwise significance tests applied.

Translation quality. The significance results support the central quality finding that metric choice changes the model ranking, especially in the literary domain. For LLaMa + FT *versus* GPT-4o + 10-shot, BLEU significantly favors LLaMa + FT in the literary domain (1/10/3), whereas COMET significantly favors GPT-4o + 10-shot in every literary direction (14/0/0). The same contrast in the news domain instead favors GPT-4o + 10-shot for both BLEU (15/1/2) and COMET (18/0/0), which strengthens the claim that GPT-4o is more uniformly dominant in news than in literary translation. The contrast between NMT + FT and GPT-4o + 10-shot gives the same interpretation: COMET significantly favors GPT in both domains (14/0/0 literary; 18/0/0 news), while BLEU is not significant in the literary domain (5/6/3) but significantly favors GPT in news (14/2/2). Thus, the statistical tests support the claim that the literary domain BLEU narrative is more favorable to fine-tuned open-source systems, whereas COMET and the news domain results more consistently support GPT-4o few-shot.

The within-family quality results also support the fine-tuning and prompting conclusions, but with domain-specific differences. In the literary domain, NMT + FT significantly improves

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table H1. Cross-direction aggregate significance for the literary domain. Each row summarizes a System A-System B contrast over all available translation directions. **B** > **A**: number of directions in which System B is significantly higher than System A, **A** > **B**: System A is significantly higher than System B, **Tie**: not significantly different, at the per-direction Holm level. **Aggregate Sig.**: two-sided binomial sign test on the **B** > **A**/**A** > **B** counts (H_0: $B > A$ count $\sim Bin((B > A)+(A > B), 0.5)$; ties excluded). **Aggregate Trend**: ↑ : System B wins more than System A and the sign test is significant; ↓ : the reverse; ≈ : sign test not significant or $B > A = A > B$. Higher is better for BLEU, COMET, and MTLD; USW and NM are descriptive (literalness). Stars: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Group	System A	System B	Metric	Trend	Aggregate sig.	$B > A$	$A > B$	Tie
Within family	NMT	NMT + FT	BLEU	↑	$p < 0.001^{***}$	14	0	0
			COMET	↑	$p < 0.001^{***}$	14	0	0
			MTLD	≈	n.s.	3	4	7
			USW	↑	$p < 0.05^{*}$	11	2	1
			NM	↑	$p < 0.05^{*}$	10	1	3
	GPT-4o	GPT-4o + 10-shot	BLEU	↑	$p < 0.01^{**}$	13	1	0
			COMET	↑	$p < 0.001^{***}$	13	0	1
			MTLD	≈	n.s.	5	0	9
			USW	↑	$p < 0.001^{***}$	13	0	1
			NM	↑	$p < 0.01^{**}$	8	0	6
	LLaMa	LLaMa + 10-shot	BLEU	↑	$p < 0.001^{***}$	14	0	0
			COMET	↑	$p < 0.001^{***}$	14	0	0
			MTLD	≈	n.s.	0	2	12
			USW	↑	$p < 0.001^{***}$	12	0	2
			NM	↑	$p < 0.01^{**}$	11	1	2
	LLaMa	LLaMa + FT	BLEU	↑	$p < 0.001^{***}$	14	0	0
			COMET	↑	$p < 0.001^{***}$	13	0	1
			MTLD	↓	$p < 0.01^{**}$	0	10	4
			USW	↑	$p < 0.001^{***}$	14	0	0
			NM	↑	$p < 0.01^{**}$	13	1	0
	LLaMa + 10-shot	LLaMa + FT	BLEU	↑	$p < 0.001^{***}$	14	0	0
			COMET	≈	n.s.	5	3	6
			MTLD	↓	$p < 0.001^{***}$	0	12	2
			USW	↑	$p < 0.001^{***}$	14	0	0
			NM	↑	$p < 0.001^{***}$	11	0	3
Cross family	NMT + FT	GPT-4o + 10-shot	BLEU	≈	n.s.	5	6	3
			COMET	↑	$p < 0.001^{***}$	14	0	0
			MTLD	↑	$p < 0.01^{**}$	13	1	0
			USW	≈	n.s.	5	3	6
			NM	≈	n.s.	3	5	6

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table H1. Continued

Group	System A	System B	Metric	Trend	Aggregate sig.	$B > A$	$A > B$	Tie
	NMT + FT	LLaMa + FT	BLEU	≈	n.s.	8	2	4
			COMET	↑	$p < 0.01^{**}$	9	0	5
			MTLD	↓	$p < 0.01^{**}$	0	10	4
			USW	↑	$p < 0.001^{***}$	14	0	0
			NM	↑	$p < 0.05^{*}$	10	1	3
	LLaMa + FT	GPT-4o + 10-shot	BLEU	↓	$p < 0.05^{*}$	1	10	3
			COMET	↑	$p < 0.001^{***}$	14	0	0
			MTLD	↑	$p < 0.001^{***}$	14	0	0
			USW	↓	$p < 0.001^{***}$	0	11	3
			NM	↓	$p < 0.01^{**}$	0	10	4

over pre-trained for both BLEU and COMET (14/0/0 for both), and LLaMa + FT significantly improves over zero-shot LLaMa for BLEU (14/0/0) and COMET (13/0/1). These results confirm that fine-tuning is a strong quality lever for open-source systems in the literary domain. In news, however, the NMT versus NMT + FT contrast is not significant at the aggregate level for BLEU or COMET despite more wins than losses, which challenges the claim that NMT fine-tuning improves news quality: the direction-specific pattern is positive but not strong enough to survive the aggregate sign test. For LLaMa, fine-tuning remains significant in both domains for BLEU and COMET, although the comparison between few-shot and fine-tuned LLaMa is significant for BLEU but not COMET in both domains, indicating that the difference between few-shot and fine-tuned LLaMa is metric-dependent. The observation that few-shot prompting yields minimal quality improvements for GPT-4o in the news domain due to a saturation effect in pre-training data is also corroborated in the significance tests, revealing no statistically significant difference in BLEU or COMET between zero-shot GPT-4o and GPT-4o + 10-shot.

Lexical diversity. The aggregate MTLD tests support the claim that GPT few-shot is the strongest lexical diversity configuration. Against fine-tuned NMT, GPT-4o + 10-shot is significantly higher in MTLD in both domains (13/1/0 literary; 16/1/1 news), and against fine-tuned LLaMa it is higher in every direction in both domains (14/0/0 literary; 18/0/0 news). These results strongly support the system-level ranking that GPT few-shot is lexically richer than the fine-tuned open-source alternatives. However, the within-family GPT-4o versus GPT-4o + 10-shot MTLD result is significant only in news (14/0/4) and not in literary (5/0/9). Therefore, the significance tests support a cautious version of the prompting claim: few-shot prompting clearly increases GPT lexical diversity in news and places GPT few-shot above the competing families in both domains, but the literary-domain MTLD increase over zero-shot GPT is not aggregate-significant.

The tests strongly support the LLaMa fine-tuning trade-off. In both domains, fine-tuned LLaMa significantly improves quality relative to zero-shot LLaMa, but MTLD significantly decreases (0/10/4 literary; 1/11/6 news). The same decrease appears when fine-tuned LLaMa is compared with LLaMa + 10-shot (0/12/2 literary; 0/9/9 news). These results directly support the claim that LLaMa fine-tuning narrows lexical variation even while improving BLEU and COMET. By contrast, NMT fine-tuning does not show an aggregate-significant MTLD change in either domain, which supports treating NMT as more lexically stable than LLaMa rather than as a system whose fine-tuning reliably increases lexical diversity.

Translation literalness. The USW and NM significance results support the claim that adaptation generally shifts outputs toward less literal translations, with fine-tuned LLaMa showing the

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table H2. Cross-direction aggregate significance for the news domain

Group	System A	System B	Metric	Trend	Aggregate sig.	$B > A$	$A > B$	Tie
Within family	NMT	NMT + FT	BLEU	≈	n.s.	10	3	5
			COMET	≈	n.s.	11	3	4
			MTLD	≈	n.s.	6	3	9
			USW	≈	n.s.	10	3	5
			NM	≈	n.s.	10	3	5
	GPT-4o	GPT-4o + 10-shot	BLEU	≈	n.s.	3	7	8
			COMET	≈	n.s.	5	0	13
			MTLD	↑	$p < 0.001^{***}$	14	0	4
			USW	↑	$p < 0.001^{***}$	13	0	5
			NM	↑	$p < 0.001^{***}$	15	0	3
	LLaMa	LLaMa + 10-shot	BLEU	↑	$p < 0.01^{**}$	13	1	4
			COMET	↑	$p < 0.001^{***}$	17	0	1
			MTLD	≈	n.s.	3	3	12
			USW	↑	$p < 0.05^{*}$	10	1	7
			NM	≈	n.s.	10	4	4
	LLaMa	LLaMa + FT	BLEU	↑	$p < 0.01^{**}$	15	2	1
			COMET	↑	$p < 0.001^{***}$	14	1	3
			MTLD	↓	$p < 0.01^{**}$	1	11	6
			USW	↑	$p < 0.001^{***}$	15	0	3
			NM	↑	$p < 0.05^{*}$	12	2	4
	LLaMa + 10-shot	LLaMa + FT	BLEU	↑	$p < 0.05^{*}$	13	3	2
			COMET	≈	n.s.	8	6	4
			MTLD	↓	$p < 0.01^{**}$	0	9	9
			USW	↑	$p < 0.001^{***}$	16	0	2
			NM	↑	$p < 0.05^{*}$	9	1	8
Cross family	NMT + FT	GPT-4o + 10-shot	BLEU	↑	$p < 0.01^{**}$	14	2	2
			COMET	↑	$p < 0.001^{***}$	18	0	0
			MTLD	↑	$p < 0.001^{***}$	16	1	1
			USW	≈	n.s.	4	9	5
			NM	≈	n.s.	2	5	11
	NMT + FT	LLaMa + FT	BLEU	≈	n.s.	10	7	1
			COMET	↑	$p < 0.01^{**}$	14	2	2
			MTLD	≈	n.s.	2	9	7
			USW	↑	$p < 0.05^{*}$	14	3	1
			NM	↑	$p < 0.01^{**}$	13	1	4

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.

Table H2. Continued

Group	System A	System B	Metric	Trend	Aggregate sig.	$B > A$	$A > B$	Tie
	LLaMa + FT	GPT-4o + 10-shot	BLEU	↑	$p < 0.001^{***}$	15	1	2
			COMET	↑	$p < 0.001^{***}$	18	0	0
			MTLD	↑	$p < 0.001^{***}$	18	0	0
			USW	↓	$p < 0.001^{***}$	0	18	0
			NM	↓	$p < 0.01^{**}$	1	13	4

clearest effect. In the literary domain, fine-tuned NMT, GPT-4o + 10-shot, LLaMa + 10-shot, and fine-tuned LLaMa all show significant increases over their corresponding baselines for USW and NM. The LLaMa versus LLaMa + FT contrast is especially strong: USW increases in all directions (14/0/0) and NM in nearly all directions (13/1/0). The news domain confirms this pattern for GPT-4o + 10-shot and fine-tuned LLaMa with significant increases in USW and NM, but for fine-tuned NMT literalness contrasts in news are not aggregate-significant despite positive win counts. Thus, the statistical evidence supports the broad adaptation-to-less-literalness claim, but the NMT news case should be described as a tendency rather than a robust aggregate effect.

The cross-family literalness results also support the distinction between lexical diversity and literalness. GPT-4o + 10-shot is lexically richer than LLaMa + FT by MTLD, yet LLaMa + FT is significantly less literal by USW and NM in both domains (USW 0/11/3 and NM 0/10/4 in literary; USW 0/18/0 and NM 1/13/4 in news). By contrast, fine-tuned NMT is not significantly less literal than GPT-4o + 10-shot at the aggregate level in either domain. This means that the claim that fine-tuned open-source models are less literal than GPT should be stated more strongly for fine-tuned LLaMa, and more cautiously for fine-tuned NMT.

Cite this article: Yirmibeşoğlu Balal Z and Güngör T (2026). Lexical diversity and translation literalness in machine translation: a comparative study of pretrained, fine-tuned, and few-shot approaches. *Natural Language Processing* **32**, 589–637. https://doi.org/10.1017/nlp.2026.10041

Downloaded from https://www.cambridge.org/core. 01 Nov 2026 at 00:31:10, subject to the Cambridge Core terms of use.