Factors Affecting Translation Quality in In-context Learning for Multilingual Medical Domain

Jonathan Mutal, Raphael Rubino, Pierrette Bouillon
TIM/FTI, University of Geneva
1205 Geneva, Switzerland
firstname.surname@unige.ch

Abstract

Multilingual machine translation in the medical domain presents critical challenges due to limited parallel data, domain-specific terminology, and the high stakes associated with translation accuracy. In this paper, we explore the potential of in-context learning (ICL) with general-purpose large language models (LLMs) as an alternative to fine-tuning. Focusing on the medical domain and low-resource languages, we evaluate an instruction-tuned LLM on a translation task across 16 languages. We address four research questions centered on prompt design, examining the impact of the number of examples, the domain and register of examples, and the example selection strategy. Our results show that prompting with one to three examples from the same register and domain as the test input leads to the largest improvements in translation quality, as measured by automatic metrics, while translation quality gains plateau with an increased number of examples. Furthermore, we find that example selection methods – lexical and embedding based – do not yield significant benefits over random selection if the register of selected examples does not match that of the test input.

1 Introduction

Multilingual communication in clinical settings is often hindered by the lack of quality translation tools for low-resource languages (Zappatore and Ruggieri, 2024). Building machine translation (MT) systems in the medical domain is challenging: parallel corpora is scarce and mistakes can lead to disastrous outcome (Chan et al., 2024). This challenge is intensified when translating into or from low-resource languages (Phan et al., 2023). Traditional neural MT models require supervised training on domain-specific data, which is not feasible for many low-resource language pairs. On the other hand, large language models (LLMs) possess broad world knowledge through pre-training and can be instructed to perform various tasks. Yet, general-purpose LLMs, when translating only with instructions (in zero-shot setting), often fail to produce adequate translations for specialized domains (Neves et al., 2024; Hu et al., 2024).

In-Context Learning (ICL) offers a way to guide LLMs at inference time by providing a few input-output pairs examples as part of the prompt (Brown et al., 2020). Unlike fine-tuning, ICL does not update model parameters; instead, the model learns from examples on the fly. This approach has gained popularity for low-resource scenarios (Zebaze et al., 2025), since only a handful of examples (as few as 1–5) can significantly improve performances on a given task. Prior studies have explored various strategies to optimize ICL for MT (Vilar et al., 2023). For instance, selection of examples that are similar to the test input yields better translation output. Similarity can be defined lexically (word overlap) or semantically (vector distance), and there is ongoing debate on which is more effective (Zebaze et al., 2025). Recent work has also examined the impact of the number of examples: some found that using up to 5 examples is beneficial (Zhu et al., 2024a), while others observed improvements up to 8 examples before performance plateau (Zhu et al., 2024b). Additionally, domain match is believed to be important: examples from the same domain as the task can guide the model's lexical and stylistic choices (Agrawal et al., 2023; Aycock and Bawden, 2024). However, to the best of our knowledge, ICL for MT in the medical domain is yet to be explored.

Our work aims to provide an evaluation of in-context learning for low-resource medical text translation. In particular, our goal is to quantify the impact of three key factors that may influence translation quality: the number of in-context examples, the register of those examples, and the strategy used to select them. We evaluate how each factor affects translation quality using two automatic metrics, namely ChrF++ (Popović, 2015)

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 161–179
November 8–9, 2025.

and COMET (Rei et al., 2020). More precisely, we seek answers to the following research questions (RQ1, RQ2, RQ3 and RQ4):

RQ1: Effect of Number of Examples – Does increasing the number of in-context examples improve translation quality?

RQ2: Effect of Register – Do examples with matching registers yield better translations than mismatched ones?

RQ3: Effect of Selection Strategy – Does semantic similarity (content-based) versus lexical similarity (form-based) selection of examples impact translation quality?

RQ4: Effect of Linguistic Characteristics - Which linguistic characteristics of the in-context examples (corpus- and prompt-level) most strongly influence translation quality?

Our contributions include: (1) an extensive empirical evaluation on 16 diverse languages (covering African, European and Asian languages and dialects) in a medical setting (Section 4), (2) a statistical analysis of the different factors, individually and in combination, that contribute to translation quality (3) a linguistic analysis of how corpus and prompts characteristics affect translation quality measured by automatic metrics (Section 5).

2 Related Work

Low-resource MT A low-resource language is typically defined as a language for which limited annotated data, such as parallel corpora or monolingual text, is available for training data-driven NLP models. The lack of resources may apply to text quantity, domain coverage, or availability of evaluation benchmarks (Joshi et al., 2020). In MT research, a low-resource language pair refers to a translation direction where parallel corpora are insufficient to train reliable MT systems. This limitation may reflect the absolute size (e.g., <1M sentence pairs), the domain (e.g., biomedical), or the bilingual coverage (Koehn and Knowles, 2017). In (bio)medical machine translation, even high-resource languages can become low-resource in-domain, due to the scarcity of domain-specific aligned corpora (e.g., medical records, Cochrane reviews, or medical dialogues) (Neves et al., 2024).

ICL for MT Early work on prompting LLMs for MT showed that models like GPT-3 can perform translation tasks without fine-tuning, particularly for high-resource languages (Brown et al., 2020). However, for low-resource languages and specialized domains, zero-shot performance is often weak, motivating research into few-shot prompting techniques (Hendy et al., 2023). One factor shown to influence translation quality is the number of examples. For instance, Peng et al. (2023) found improvements up to about five examples, while Zhu et al. (2024c) reported gains up to eight examples before saturation (translation quality scores plateau). These differences suggest that task- or model-specific characteristics can affect translation quality. This research direction allows us to answer RQ1.

Few-shot selection The example selection strategy for ICL is another impacting factor for MT, which has produced mixed results. Vilar et al. (2023) did not observe significant differences between random examples and lexically similar examples in some setups. In contrast, Zebaze et al. (2025) found that a semantic or lexical selection of examples based on the similarity with the source text to be translated improves translation quality. Moslem et al. (2023) proposed a more fine-grained approach, identifying which source words contribute most to guide example selection. More recently, Zebaze et al. (2025) showed that a small number of similar examples can yield large gains in translation quality for low-resource languages, even if the impact is limited for high-resource pairs where the LLM is already strong. We contribute to this debate by comparing form-based and meaning-based retrieval (answering RQ3), using BM25 (Robertson and Zaragoza, 2009) and LASER (Artetxe and Schwenk, 2019), respectively. This comparison has not been extensively explored in a multilingual medical setting.

Domain, Register and MT Domain is another influencing factor in MT quality. Farajian et al. (2017) observed that domain-matching data between training and testing leads to MT improvement, which motivates our experiments on few-shot selection from various sample pools, including in- and out-of-domain corpora. Following these ideas, Agrawal et al. (2023); Sia and Duh (2023); Aycock and Bawden (2024) showed that using ICL for MT, using in-domain data (e.g., medical or legal) as examples helps the model to produce appropriate terminology and style. In this work, we examine not only the domain but also the register (Lecorvé et al., 2023) of the texts (RQ2), as both can influence translation quality – an aspect that has received little attention in prior work. While

domain refers to the subject matter of a text, such as medicine, law, or education, register describes how language is used in a specific situation within that domain, shaped by factors like the relationship between speakers, the communication channel, and the purpose of the interaction. For example, both a doctor–patient dialogue and a medical research paper belong to the medical domain while being in different registers and thus exhibit different styles, choices in vocabulary, and overall communicative intent.

3 Methodology

3.1 Test Data

Our evaluation spans 16 languages: Albanian, Modern Standard Arabic, Moroccan Arabic, Tunisian Arabic, Dari, Farsi (Persian), Russian, Romanian, Ukrainian, English, French, Spanish, German, Polish, Czech and Tigrinya. These include low-resource languages/dialects (e.g. Moroccan, Tunisian and Tigrinya) as well as higher-resource ones (French, Spanish). We consider translation between all pairs and translation directions among these languages.

We evaluate various prompt engineering settings on three test sets in the medical domain that differ in register. These test sets are n-way parallel–each sentence translated into multiple languages–thus allowing us to assess: i) cross-linguistic variations, ii) differences in style and iii) communicative purposes within medical texts.

Cochrane is an internationally recognized source of evidence-based clinical research, providing reviews that synthesize medical studies to inform clinical practice[1]. The language used in Cochrane documents is formal, technical, and structured, making it representative of technical biomedical content. The content of this article is aimed at medical professionals, particularly researchers in the medical field.

NHS24 consists in publicly available health articles from Scotland's national telehealth service [2]. These articles are designed for the general public and provide accessible medical information, symptom explanations, and healthcare guidance. The language used in this article aims to be understandable, non-technical, and oriented toward patient comprehension, distinguishing it from more technical registers such as Cochrane. This corpus represents a patient-facing, health communication register, in a scenario where translation clarity and simplicity are critical.

Medical Dialogues is a set of medical question-and-instruction sentences from Bouillon et al. (2021). This corpus has never been released publicly, thus constitute an annotated *no-leakage* evaluation set never seen by LLMs. Sentences in this corpus are characterised by short, directive, and information-seeking utterances typical of clinician–patient interactions (e.g., asking about symptoms, giving instructions for treatment, etc.). The sentences were translated from French.[3]

An important characteristic of our medical dialogue data is that the translators were instructed to generate target-oriented translation that read as if originally written in the target language. They were also asked to take the communicative context and audience into account (e.g., patient vs. clinician) and allowed freedom of reformulation rather than adhering to the source structure. As a result, the dataset avoids many of the typical artifacts of translationese – such as literal lexical choices, unnatural word order, or oversimplification – while still maintaining terminological accuracy (Gerlach et al., 2018). An example of the translations is shown in Table 2. For the other datasets, we acknowledge the potential influence of translationese, but because both training and evaluation rely on the same language pairs (including artificial test pairs), the artifacts introduced by translationese is expected to be consistent across sets and therefore less likely to distort automatic scores (Ni et al., 2022).

3.2 Register and Domain Selection

To assess the effect of test and n-shot source mismatch, various datasets are used as sources for few-shot sampling: datasets described in Section 3.1, as well as a general-domain corpus: FLORES+ (Team et al., 2022). In our experiments, we control for all other factors (model, language pair, test sentence, prompt length, and evaluation metric) and vary only the source of the in-context examples, allow-

[1] An example can be found at https://pmc.ncbi.nlm.nih.gov/articles/PMC7045447/

[2] For instance, https://www.nhsinform.scot/healthy-living/preventing-falls/falls-and-dementia/

[3] The dataset is available in https://huggingface.co/datasets/jonathanmutal/Medical-Questionnaire-Multilingual-Translation

ing us to compare matched vs. mismatched domain and register. We follow the standard dataset splits for Cochrane and NHS24 (Haddow, 2015) and for FLORES+ (Team et al., 2022). For the Medical Dialogues set, we randomly extracted 1,000 segments. The number of n-way parallel segments for each dataset is shown in Table1.

Dataset	Split	#Sentences
Cochrane	Train	759
	Test	672
NHS24	Train	1200
	Test	1257
Medical Dialogues	Train	8511
	Test	1000
FLORES+	Train	997

Table 1: Number of sentences for each dataset and split.

3.3 In-Context Example Selection

To address RQ3, we consider three example selection strategies:

Random: We randomly sample n examples from the available pool of parallel data. This acts as a baseline and helps quantify variance. To ensure fairness, the same set of random examples (for a given n) is used across different languages when evaluating the number of examples. This way, any observed differences are not due to content variations across languages.

Lexical Similarity (BM25): We retrieve examples that are lexically similar to the input, using the BM25 (Robertson and Zaragoza, 2009) ranking function. BM25 scores examples based on overlapping words (with term-frequency and length normalization). This method prioritizes examples containing similar medical terms or phrases, reinforcing consistent terminology.

Semantic Similarity (LASER): We use multilingual LASER embeddings (Artetxe and Schwenk, 2019) to find examples with high cosine similarity to the input sentence. This can retrieve examples that are paraphrases or semantically related, even if they do not share keywords. The goal is to help the model generalize to similar meanings expressed using various surface forms.

For both BM25 and LASER, we retrieve the top n examples from each dataset individually. We also follow a specific ordering when placing examples in the prompt: we sort the retrieved examples by descending similarity to the input (most similar last, closest to the input). This ordering, suggested by prior work (Chitale et al., 2024), may maximize the utility of the demonstration closest to the test query.

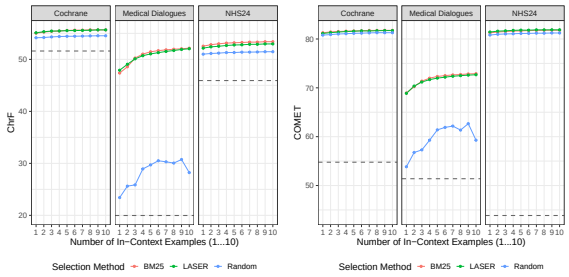


Figure 1: Effect of the number of in-context examples on ChrF scores across datasets and selection methods. ChrF performance is plotted for three datasets – Cochrane, Medical Dialogues, and NHS24– using three selection strategies: BM25 (red), LASER (green), and Random (blue). Scores are averaged across all languages. The dashed horizontal lines represent zero-shot performance.

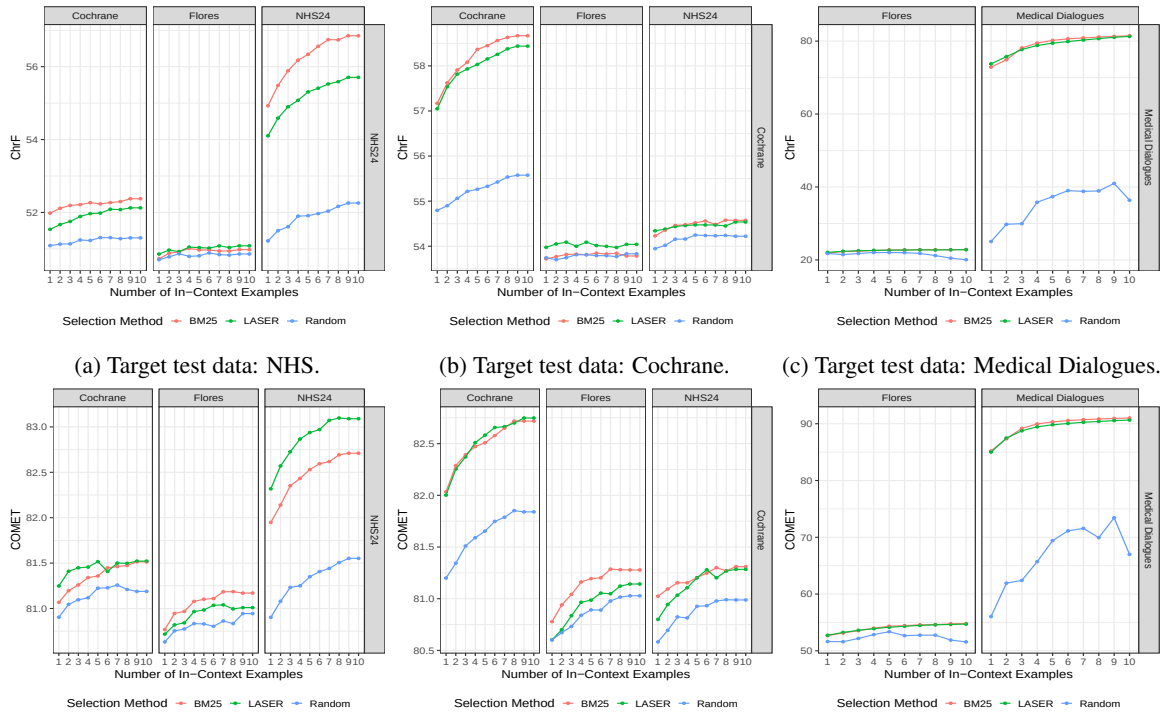
3.4 Experimental Settings

We use Mistral-7B-Instruct[4] with a fixed JSON-based prompt format to ensure systematic outputs and isolate the effect of example content on translation quality (cf. Appendix B for more details). We vary the number of in-context examples n from 0 (zero-shot, only instruction) up to 10 to address RQ1. Each configuration (defined by the number of examples n, the selection method, and the domain of the selected examples) is applied to all test sentences. For stochastic settings (random selection), we repeat each test 30 times with different random seeds and average the results. This yields robust estimates of performance by smoothing the results obtained with random sampling and allows for significance testing.

We evaluate translation quality with two automatic metrics:

- ChrF++: Character n-gram F-score, which correlates well with adequacy especially for morphologically rich languages (Popović, 2015).
- COMET: A learned metric that predicts human judgment scores using multilingual em-

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3



(a) Target test data: NHS. (b) Target test data: Cochrane. (c) Target test data: Medical Dialogues.

(d) Target test data: NHS. (e) Target test data: Cochrane. (f) Target test data: Medical Dialogues.

Figure 2: Effect of example domain on translation quality across selection methods. ChrF and COMET scores are shown for in-context examples drawn from different domains using three selection methods (BM25, LASER, RANDOM).

beddings; effective for capturing semantic adequacy (Rei et al., 2020).

All metrics are computed against reference translations. We conduct significance testing Factorial Analysis of Variance (Factorial ANOVA, Ross and Willson, 2017) to understand the effect of each factor (number of examples, register and selection strategy) on the translation quality, and also understand the effect of multiple factors on translation quality. This analysis allows us to understand, for example, the effect of the number of examples and selection strategy on the translation quality. We used eta squared (η^2) test (Adams and Conway, 2014) to quantify the effect size in analysis of variance and determine which factor has the largest effect on translation quality according to automatic metrics.

4 Effect of Factors in Translation Quality

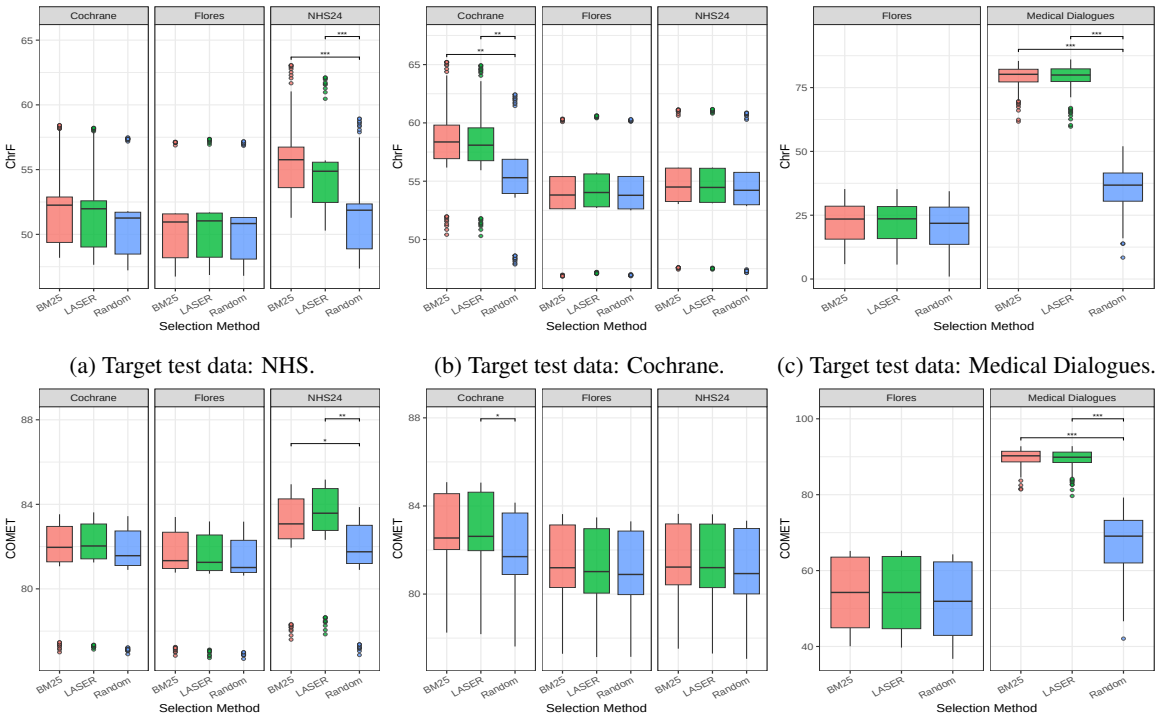
In this section, we describe the results collected during our experiments. We divide the results following the RQ1, RQ2 and RQ3 from Section 1 before comparing the effects of all factors on translation quality.

4.1 Results

Number of Examples: Figure 1 illustrates the effect of the number of examples on translation quality. We observe that increasing the number of in-context examples improves translation quality for all test sets, but differences are observed in terms of n-shot configuration.

With no translation examples in the prompt (0-shot, the dashed line in Figure 1), the LLM reaches the lowest scores on the medical dialogues test set. With just a single example, automatic scores more than doubled (from 20.01 to 48.20 ChrF using BM25). For the best selection method on this test set (medical dialogues), there are diminishing returns above 4 to 5-shot configuration.

For Cochrane and NHS24, the difference between 0-shot and n-shot is smaller based on ChrF (51.25 vs. 55.21 for Cochrane, 45.23 vs. 53.25 for NHS24). For these particular test sets, most gains are obtained with 2 to 3 examples, followed by a plateau with no significant improvement when increasing the number of shots. This may be due to the fact that the Cochrane and NHS24 datasets are publicly available and may have been seen by the LLM during pre-training, whereas this is not



(a) Target test data: NHS. (b) Target test data: Cochrane. (c) Target test data: Medical Dialogues.

(d) Target test data: NHS. (e) Target test data: Cochrane. (f) Target test data: Medical Dialogues.

Figure 3: Effect of selection method on translation quality across different test datasets. ChrF and COMET scores are shown for in-context examples drawn from different selection methods (BM25, LASER, RANDOM) on the different domain examples. Statistical significance among results is indicated by *** when $p < 0.001$, ** when $p < 0.01$ and * when $p < 0.05$.

the case for Medical Dialogues.

To answer RQ1, these results show that increasing the number of examples has an effect on translation scores measured by automatic metrics. However, a plateau is quickly reached on all test sets (max. with a 5-shot prompt), and adding more examples does not lead to significant improvements. This is observed when examples are retrieved based on their similarity with the source using BM25 and LASER.

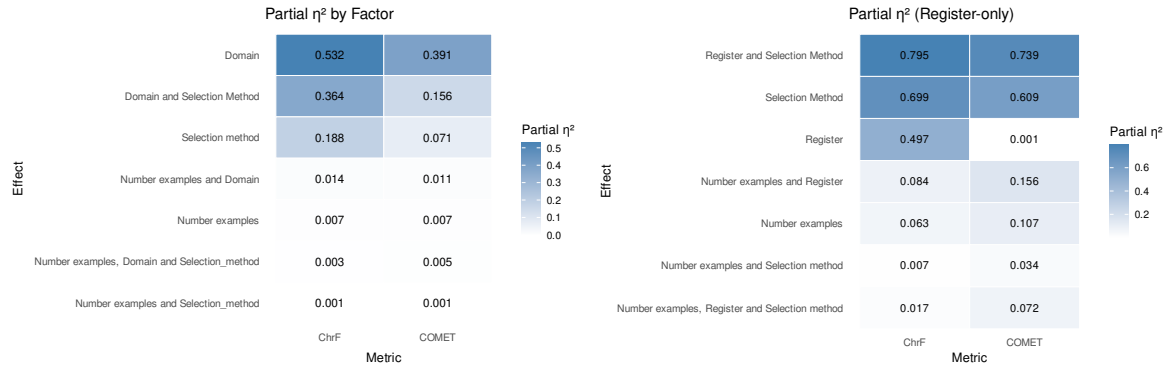
Source of Examples: Figure 2 shows translation results on our test sets when various sample pools are used to build the prompt.

Matching-domain exemplars (n-shot and test sentences sampled from the same corpus) result in significant score gains for all selection methods. The figures show that sampling from FLORES+ consistently underperforms compared to sampling from in-domain datasets. We also observe that random sampling from in-domain data outperforms a selection strategy using an out-of-domain dataset, showing that the domain of the data have a strong effect on translation quality.

Additionally, results show that matching registers yields the highest scores. However, unlike domain mismatch, we found that using a non-random selection method with Cochrane on the NHS24 test set yields higher ChrF scores compared to randomly sampling from the same register (sampling from NHS24), although the difference is not statistically significant according to COMET. This may be due to the content of the NHS24 corpus, written for patients and thus using less technical vocabulary compared to Cochrane.

To answer RQ2, the results show that register and domain have a significant effect on translation quality.

Example Selection Method: We compare lexical vs. semantic retrieval (BM25 vs. LASER, respectively) against random examples selection to address RQ3. Figure 3 illustrates the box plot for the different test data with the n-shot retrieval methods. Overall, BM25 and LASER yield nearly identically scores on automatic metrics. BM25 had a slight higher automatic scores (but not statistically significant). We found that when using BM25 for lower-resource languages seems to be more beneficial (we refer to Figure 10 for these



(a) Partial η^2 for different domains and registers.

(b) Partial η^2 for the same domain and different registers.

Figure 4: Heatmap of partial η^2 values indicating the percentage of variance explained by each factor and interaction in the model. Darker shades represent greater effect on translation quality. Register and Selection method show the highest effects, while all interactions involving Number examples contribute minimally. Note that partial eta-squared values are not additive and do not sum to 100%.

results).

BM25 and LASER achieve higher translation quality, as measured by the automatic metrics, compared to random sampling when the domain and register match those of the test data. However, when using data from a different domain or register, the selection methods do not yield significant improvements compared to random sampling. This provides further evidence of n-shot domain and register impact on translation quality, adding support to the findings for RQ2. We can therefore answer RQ3: there is an effect of selection strategy when the examples match the register. Otherwise, there is no statistical difference between BM24, LASER and random selection.

Comparing Effects: We conducted a factorial ANOVA to quantify the contribution of each factor to the variance in translation quality, as measured by ChrF and COMET, illustrated by Figure 4a.

Using common benchmarks for partial η^2 ($\sim$ .01 small, $\sim$ .06 medium, $\sim$ .14 large)[5], this analysis reveals that register and domain shows the largest effect ($\eta^2 = 0.53$), suggesting the highest variance in translation quality is associated with whether the train and test are sampled from the same dataset. The selection method ($\eta^2 = 0.18$) has also a large impact on translation quality. However, the interaction between the selection method and matching dataset effects on translation quality is higher (η^2 = .36), indicating that the selection method has a larger impact only when accompanied by matching n-shot register and domain.

[5] https://resources.nu.edu/statsresources/eta

On the other hand, increasing the number of examples in the prompt does not seem to have a strong impact ($\eta^2 = .0028$) compared to the other factors, i.e. domain and register. This supports the importance of the data source for n-shot selection. Increasing the number of examples provided as prompt to the LLM shows small additional variance on translation quality (above 1-shot).

When measuring translation quality variance within the medical domain, the main drivers are the selection method ($\eta^2 = 0.699$) and its interaction with the register ($\eta^2 = 0.795$). Number of examples is modest ($\eta^2 = 0.063$), and other interactions are small. Within the same domain, the choice of example selection strategy have a strong influence on both ChrF and COMET scores ($\eta^2 = 0.795$ and $\eta^2 = 0.739$ respectively), with its impact varying across registers. Figure 4b illustrates this values in a heatmap figure.

These results confirm our findings of the most important factor of translation quality: register and domain of the examples. To understand the causes, we carry out a lingistic evaluation in the following section.

5 Linguistic Analysis

Based on the previous results, which showed that the domain and register of the examples have the strongest impact on translation quality, we conduct a linguistic evaluation to determine which linguistic characteristics from the examples have the most influence on translation quality (RQ4). Because our goal is to assess how domain and register alignment shape translation quality, we select three corpus-

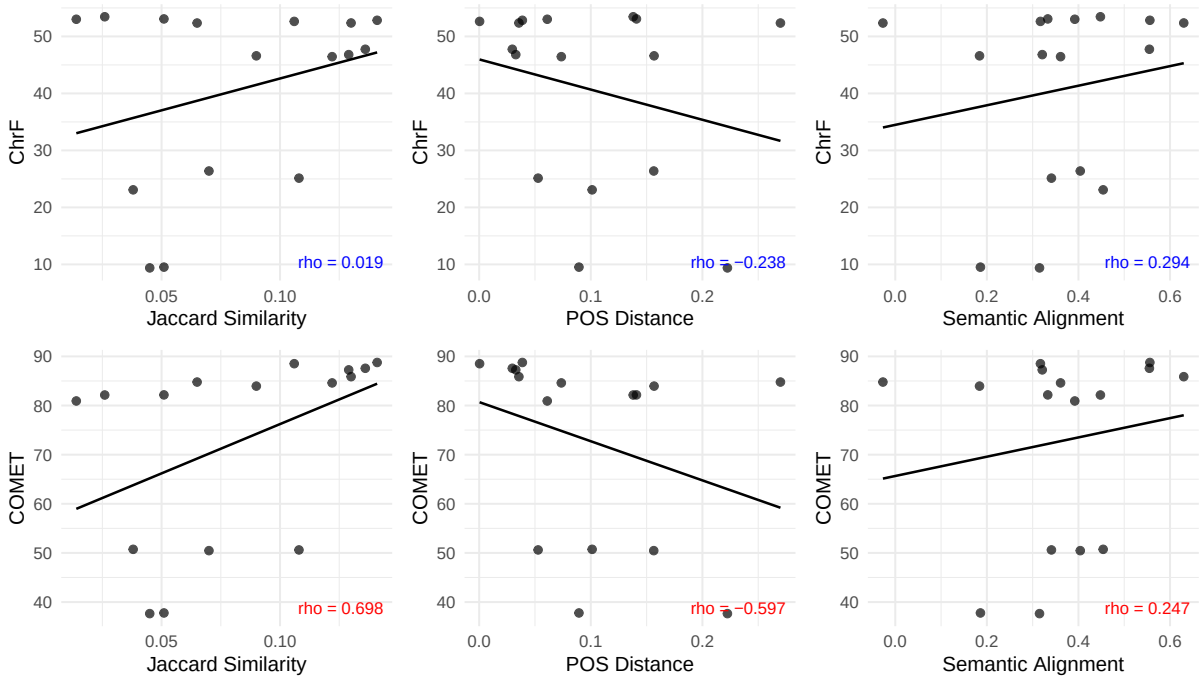


Figure 5: Scatterplots showing the relationship between corpus-level similarity metrics and average translation quality (ChrF and COMET). Each plot corresponds to one of the following corpus-level measurements: **Jaccard Similarity** (lexical overlap), **POS Distance** (cosine distance between part-of-speech distributions), and **Semantic Alignment** (average of sentence cosine similarity between corpus). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

level metrics that capture complementary aspects of similarity in texts:

- **Lexical overlap** (using (Jaccard similarity Jaccard, 1901) for corpus-level analysis and token overlap for prompt-level analysis): it measures vocabulary overlap across examples to capture whether they share key medical terms, and thus belong to the same domain and register. High overlap indicates coherence in subject matter.

- **Structural similarity** (cosine difference of part-of-speech distributions) (Liu et al., 2021): Part-of-speech distributions reflect grammatical choices; their cosine difference approximates whether examples adopt similar interpersonal stances and modes of medical communication.

- **Semantic alignment** (cosine similarity of sentence embeddings): Compares sentence embeddings to assess whether the overall meaning of the examples aligns. We used a different sentence embedding model from that used in the selection method, specifically the one proposed by (Reimers and Gurevych, 2019).

To perform this analysis, we included additional corpora for n-shot sampling to provide more data for corpus-level analysis: translations of documents related to public health and disease prevention across different languages within the European Union (ECDC, European Centre for Disease Prevention and Control, Greer, 2012); diverse datasets created during the COVID-19 period, including a set of Wikipedia documents related to health (Wikipedia Health); a database containing European Union law and other public documents generated during COVID-19[6]; and TICO-19 for non-European languages (Anastasopoulos et al., 2020). Finally, we included a general-domain corpus, Tatoeba[7]. While several of these texts share the same domain (medical), they differ in register, ranging from policy documents (EU public documents) to encyclopedic health texts and public health advisories. All corpora were extracted using the OPUS platform (Tiedemann, 2012).

We first examine corpus-level characteristics to understand how to select sample pools for ICL and to determine which aspects of domain and register are the most impactful when choosing a sample

[6]Extracted from https://elrc-share.eu/

[7]https://tatoeba.org/en/

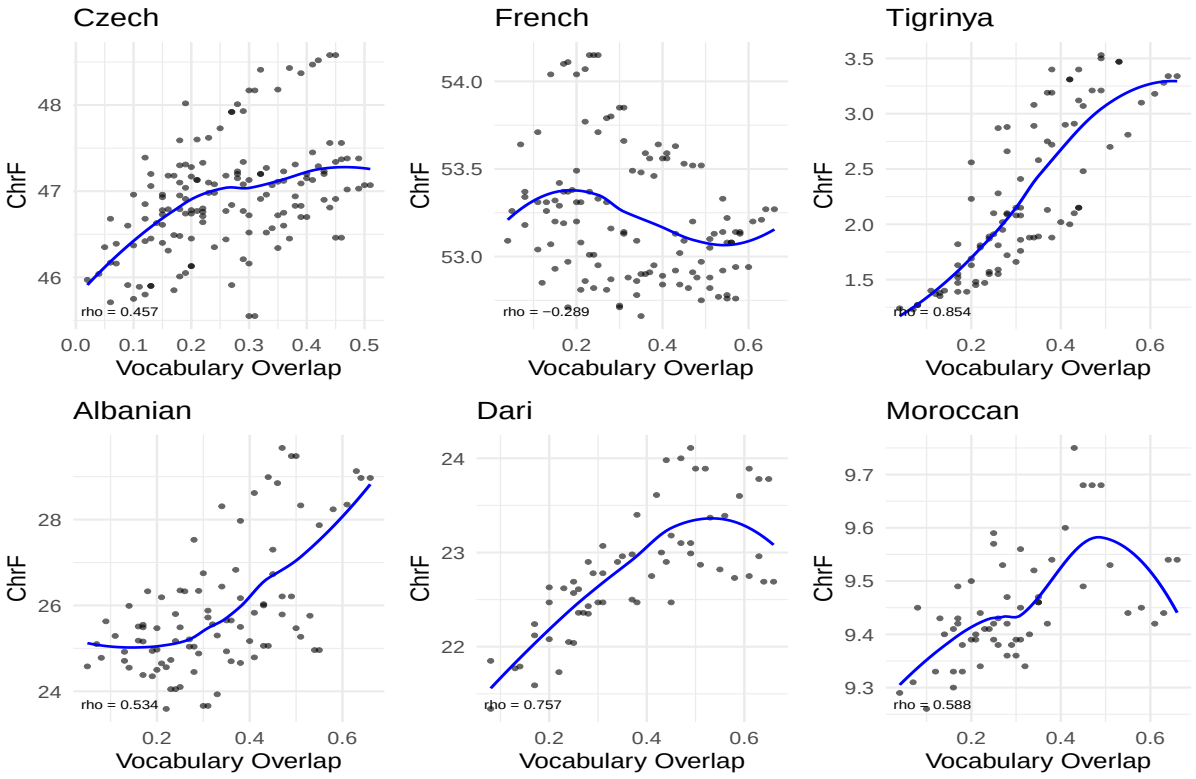


Figure 6: Scatterplots showing the relationship between prompt-level similarity metrics and average translation quality. Each plot corresponds to one of the following prompt-level measurement for **Vocabulary Overlap** (lexical overlap). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

pool. We then analysed prompt-level examples[8] to evaluate the effect of their linguistic characteristics on translation quality. To identify which aspects of register matter most, we examined the relationship between linguistic features and evaluation scores, checking how strongly each feature–such as vocabulary overlap, grammatical similarity, and semantic similarity–correlated with ChrF and COMET using Spearman rank correlation.

We selected language pairs in which POS tagging and embedding-based metrics for the source language were supported and for which the neccesary corpora were available. Specifically, we used English paired with six languages: two high-resource languages (French and Czech) and four low-resource languages (Tigrinya, Albanian, Moroccan Arabic, and Dari), selected based on the availability of medical corpora.[9]

[8]When multiple examples were provided for in-context learning, we calculated the mean; maximum scores were also tested but showed lower correlation with translation quality according to Pearson's ρ.

[9]According to the OPUS platform, English–Tigrinya has 6,142 sentences in the COVID medical domain; English–Dari has 3,071; English–Albanian has 389; and English–Moroccan Arabic has none.

In the next section, we describe the results collected during our experiences to answer RQ4, divided by corpus and prompt levels of analysis.

5.1 Results

Corpus-level Analysis At the corpus level, the similarity measures (lexical, syntactic, semantic) are the same for all sentences within a given sampling and test dataset, as they are calculated over the full example set, while translation quality varies between prompts. To avoid inflating the number of independent observations, we averaged the translation quality scores for all sentences in the same sampling and test dataset configuration and used these aggregated values in the analysis. This ensures that each configuration is counted once, and the results reflect real differences between configurations rather than repetition of identical feature values.

Figure 5 illustrates the corpus-level analysis between the translation automatic scores and the corpus-level metrics for lexical overlap. The results show that lexical overlap strongly correlates with COMET ($\rho = 0.698$), but not with ChrF, suggest-

ing that COMET is more sensitive to lexical content alignment. POS distance shows negative correlations with both metrics, especially with COMET, indicating that structural divergence between examples and test sets degrades ICL performance. Semantic alignment correlates moderately with both ChrF ($\rho = 0.294$) and COMET ($\rho = 0.247$), confirming that semantically coherent prompts are beneficial, although their predictive power is lower than the lexical alignment for COMET.

Prompt-level Analysis To assess the effect of samples register and domain from the prompt on translation quality, we first calculated the Spearman's ρ correlation between the translation quality scores and the linguistic features between the samples and the input sentence – vocabulary overlap, grammatical similarity, and semantic similarity.

Figures 6, 11, 12 show that translation quality–measured by ChrF and COMET–correlates with lexical, syntactic, and semantic similarity between the input and the selected examples. Spearman correlations indicate that low-resource languages such as Tigrinya, Dari, Moroccan Arabic, and Albanian exhibit the strongest correlations across all three similarity types, while higher-resource languages display more selective patterns. The ANOVA analysis, which includes the number of examples and the selection method as fixed effects, confirm these trends, with semantic similarity often producing the largest effect in translation quality for low-resource languages, and syntactic similarity dominating in Czech. η^2 analysis further reveals the unique contribution of each feature: semantic similarity explains the largest share of variance in most low-resource languages (e.g., 4–10% in Albanian and Moroccan Arabic), whereas in French lexical similarity accounts for 13–15% and in Czech syntactic similarity explains up to 21.7% of variance in ChrF. Together, these results show that the relative importance of lexical, syntactic, and semantic alignment is language-dependent.[10]

6 Conclusions

This study shows that, in multilingual medical machine translation, the domain and register of in-context examples are the most influential factors affecting translation quality. Partial η^2 analysis confirms that aligning the n-shot register and domain with the test input yields substantially greater improvements than increasing the number of examples. In practice, a small, well-chosen set of domain-relevant shots often yields higher translation quality scores than a larger set of examples sampled from other domains or registers.

Sentence-level analysis of lexical, syntactic, and semantic similarity confirms that the most predictive features vary by language. In low-resource language pairs, all three similarity types correlate strongly with translation quality, while in higher-resource languages pairs such as English to Czech and French, syntactic and semantic similarity dominate. Semantic similarity is the most consistent predictor across languages.

These results suggest that prompt engineering for ICL should prioritise register and domain alignment, and adapt exemplar selection criteria to the characteristics of the language pair rather than applying the same similarity heuristics.

Acknowledgements

This work is part of the PROPICTO project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005). We would also like to thank the three reviewers for their careful suggestions, which helped improve this work.

Limitations

This study has several limitations. First, all experiments were conducted using a single LLM, which constrains the generalisability of the findings to other model families, training paradigms and sizes. We hypothesize that larger models could reach better translation quality, which we leave for future work. Second, the linguistic similarity features—lexical, syntactic, and semantic—were computed using specific operationalisations (e.g., Jaccard similarity, POS distribution cosine distance, sentence embedding cosine similarity). They represent only one way of quantifying similarity, and alternative feature definitions or embeddings might yield different rankings of predictive importance. Moreover, corpus-level features were constant within each configuration, which required aggregation to avoid artificially statistical significance; this design limits the granularity of the corpus-level analysis. The linguistic evaluation was limited to a fixed set of high- and low-resource languages in the medical domain, meaning that

[10] Type–token ratio was negatively correlated with both metrics in nearly all languages, suggesting that higher lexical diversity in prompts tends to reduce translation quality.

results may not generalise to other languages. Finally, while ChrF and COMET provide complementary perspectives on translation quality, incorporating human evaluation for adequacy and fluency would strengthen the validity of the results. Furthermore, the evidence gathered in this work provides practical insights into the factors influencing translation quality as measured by automatic metrics. However, these findings do not indicate whether the translations are sufficiently accurate for practical use without introducing potential risks. Future work will involve evaluating clinical risks, following the approach of Mehandru et al. (2023).

References

Marc A. Adams and Terry L. Conway. 2014. *Eta Squared*, pages 1965–1966. Springer Netherlands, Dordrecht.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. 2021. A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 135–142, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, page 1877–1901. Curran Associates, Inc.

Brittany M. C. Chan, Jeanine Suurmond, Julia C. M. Van Weert, and Barbara C. Schouten. 2024. Uncovering communication strategies used in language-discordant consultations with people who are migrants: Qualitative interviews with healthcare providers. *Health Expectations*, 27(1):e13949.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Johanna Gerlach, Hervé SPECHBACH, and Pierrette Bouillon. 2018. Creating an online translation platform to build target language resources for a medical phraselator. In *Proceedings of the 40th edition of Translating and the Computer Conference (TC40)*, pages 60–65. AsLing, The International Association for Advancement in Language Technology, London (UK).

Scott L Greer. 2012. The european centre for disease prevention and control: hub or hollow core ? *Journal of health politics, policy and law*, 3737(6)(1):1001—-1030.

Barry Haddow. 2015. HimL (health in my language). In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey. European Association for Machine Translation.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles (in French)*, 37 (142):547–579.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, Delphine Battistelli, and Nicolas Béchet. 2023. Towards the automatic processing of language registers: Semi-supervisedly built corpus and classifier for french. In *Computational Linguistics and Intelligent Text Processing*, pages 480–492, Cham. Springer Nature Switzerland.

Zeyang Liu, Ke Zhou, Jiaxin Mao, and Max L. Wilson. 2021. Posscore: A simple yet effective evaluation of conversational search with part of speech labelling. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1119–1129, New York, NY, USA. Association for Computing Machinery.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Long Phan, Tai Dang, Hieu Tran, Trieu H. Trinh, Vy Phan, Lam D. Chau, and Minh-Thang Luong. 2023. Enriching biomedical knowledge for low-resource language through large-scale translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3131–3142, Dubrovnik, Croatia. Association for Computational Linguistics.

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Amanda Ross and Victor L. Willson. 2017. *Factorial Anova*, pages 25–29. SensePublishers, Rotterdam.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht,

Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Marco Zappatore and Gilda Ruggieri. 2024. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, 84:101582.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

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

Yuanyi Zhu, Maria Liakata, and Giovanni Montana. 2024c. A multi-task transformer model for fine-grained labelling of chest X-ray reports. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 862–875, Torino, Italia. ELRA and ICCL.

A Example Translation Corpora

Table 2 illustrates a typical sentence from our medical dialogue dataset in French, English, and Spanish. As seen in the example, the translations are target-oriented and adapted to the communicative context: the English version uses an idiomatic rendering ("ringing noise in your ears"), while the Spanish version employs an equivalent ("zumbidos") rather than a literal calque of the French source term. This reflects the dataset's design guidelines, which emphasized the audience awareness and freedom of reformulation.

Language	Sentence
French	Pendant combien de jours avez-vous pris des médicaments contre les acouphènes ?
English	How many days did you take medicine to help with the ringing noise in your ears for?
Spanish	¿Durante cuántos días ha estado tomando medicamentos contra los zumbidos?

Table 2: Example of a medical dialogue sentence in three languages.

B Settings

B.1 Model Prompt and Design

```
Can you translate from English to French ?

Return the result in JSON format with the following schema:

{{
  "translation": {{
    "type": "string"
  }}
}}

Generate the translation for the text that appears after <<<>>>.
Do not provide explanations or additional comments. You can return only one variation.

###
Here are some examples:

English: I will give you a prescription for cortisone
French JSON: {'translation': 'je vais vous prescrire de la cortisone'}

English: I will give you a prescription for medicine with cortisone in it
French JSON: {'translation': 'je vais vous prescrire des médicaments à base de cortisone'}

English: I will give you a prescription for steroids
French JSON: {'translation': 'je vais vous prescrire des stéroïdes'}

English: I will give you a prescription for a cream
French JSON: {'translation': 'je vais vous prescrire une crème'}

###
<<<
English: I will give you a prescription for a cortisone cream. Cortisone is a steroid that helps stop swelling.
>>>
```

Figure 7: Prompt structure for in-context learning, illustrated for English-to-French text translation. The prompt provides an instruction with output schema, a few example input-output pairs in JSON format, and then the test input demarcated by special tokens.

We use Mistral-7B-Instruct v0.3, a 7-billion parameter decoder-only LLM, as the backbone. This model was chosen because at the time of experimentation it was one of the stronger openly-available instruction-tuned models. Notably, Mistral-Instruct is predominantly trained on English and lacks dedicated support for many of our languages (e.g. Tigrinya), making it a good stress-test for ICL. We access the model via HuggingFace Transformers, running in half-precision (fp16) with FlashAttention optimization for efficiency. Generation is done greedily (no sampling) to ensure deterministic outputs for a given prompt.

We construct a prompt template that includes an instruction section, a few example translation pairs, and then the input to translate. The instruction defines the task (e.g., "Translate from language X to language Y and output in a JSON format"). We enforce a JSON output schema to ensure the model's output is structured correctly. An example prompt (for English-to-French translation) is shown in Figure 7. The prompt begins with a task description and output schema specification (the schema indicates that the output should be a JSON with a "translation" field containing a string). It then says: "Here are some examples:" followed by N example pairs. Each example is formatted as:

[source language]: [source text example]

[target language] JSON: "translation": "[target text example]"

After listing the N examples, the prompt has a separator and then the actual input to be translated, marked clearly (e.g., by <<< >>>). The model is expected to produce only the JSON translation for the input without additional commentary. We found that including the language names (as in the figure) helps the model produce the output in the correct language, especially since the model is multi-lingual only through prompting. This prompt format was kept consistent in all experiments to focus on the content of examples rather than prompt wording.

C Effect of Factors on Translation Quality

Figures 9, 8 and 10 show the detailed results for the effect of the number of examples and the selection method across test sets by langauge. Each subfigure presents ChrF scores for examples drawn from different registers (BM25, LASER, RANDOM). Statistical significance between registers is indicated in the plots. These results complement the main findings in Section 4.1, providing per-dataset and per-method breakdowns that were summarised in the main text.

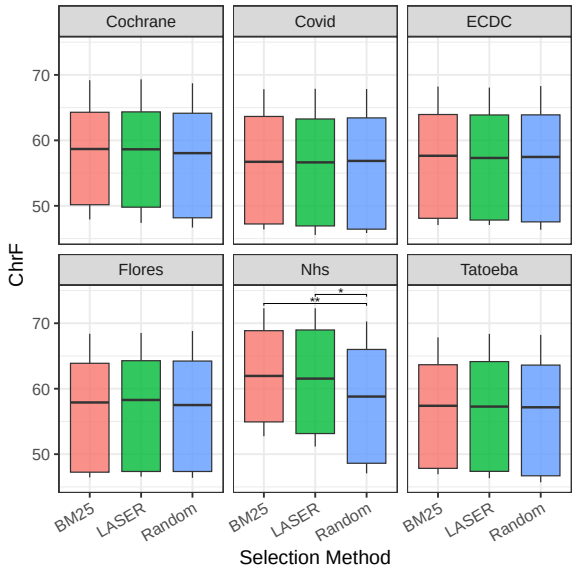


Figure 8: Effect of selection method on translation quality across different test datasets tested in NHS24. ChrF and COMET scores are shown for in-context examples drawn from different selection methods (BM25, LASER, RANDOM) on the different domain examples. Statistical significance among results is indicated by *** when $p < 0.001$, ** when $p < 0.01$ and * when $p < 0.05$.

D Linguistic Evaluation

Figures 11 and 12 present the full scatterplots for the relationship between prompt-level similarity metrics and average translation quality. Each figure corresponds to one similarity feature:

- Figure 11: **POS Distance** (cosine distance between part-of-speech distributions).
- Figure 12: **Semantic Alignment** (cosine similarity between sentence embeddings).

Spearman's ρ is shown for each plot, indicating both the strength and direction of the correlation. These figures provide the complete visual evidence underlying the correlation values reported in Section 5.

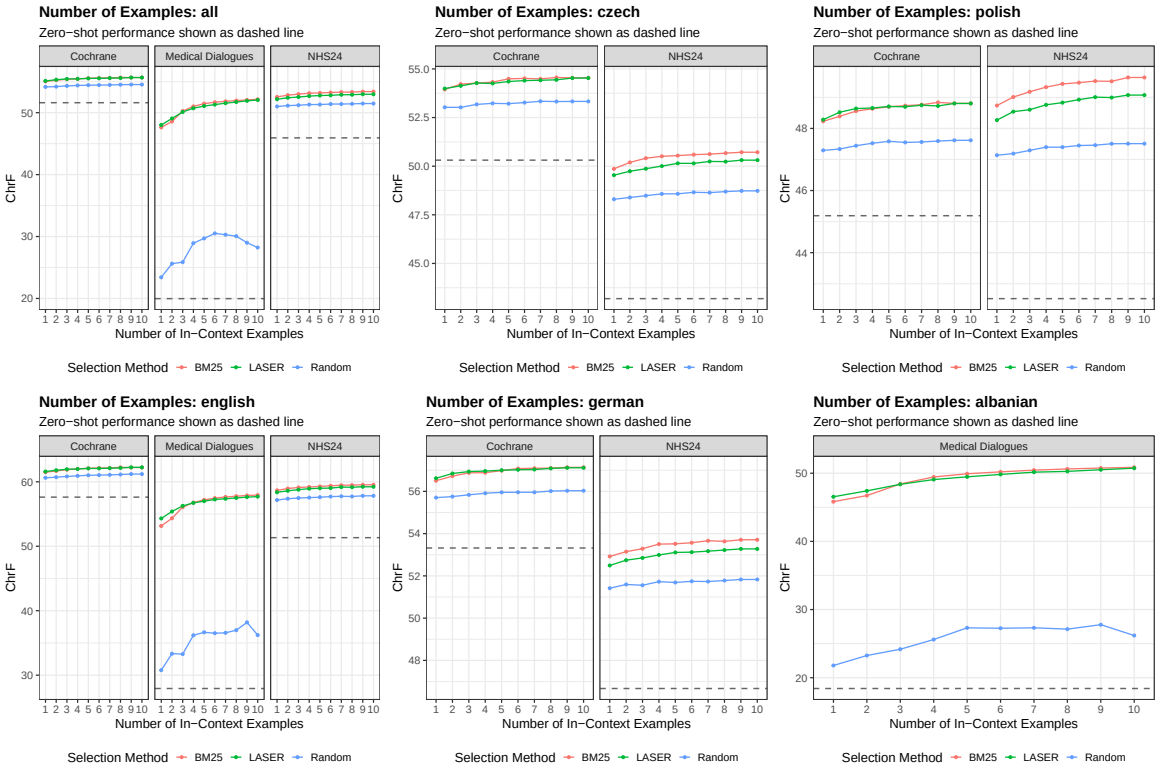


Figure 9: Effect of number of examples and selection method per test data. ChrF scores are shown for in-context examples drawn from different registers using three selection methods (BM25, LASER, RANDOM). Statistical significance between registers is indicated.

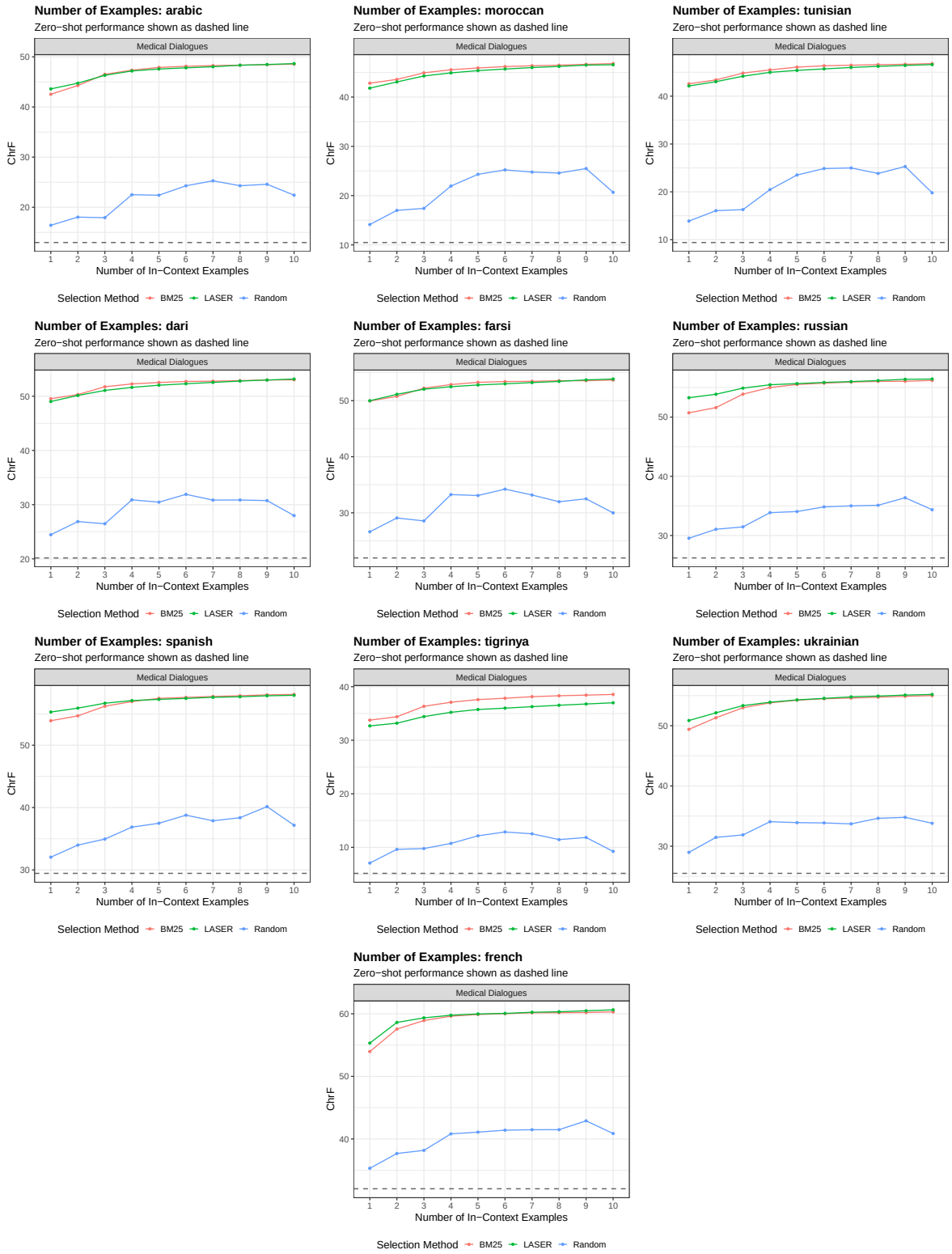


Figure 10: Effect of number of examples and selection method per test data. ChrF scores are shown for in-context examples drawn from different registers using three selection methods (BM25, LASER, RANDOM). Statistical significance between registers is indicated.

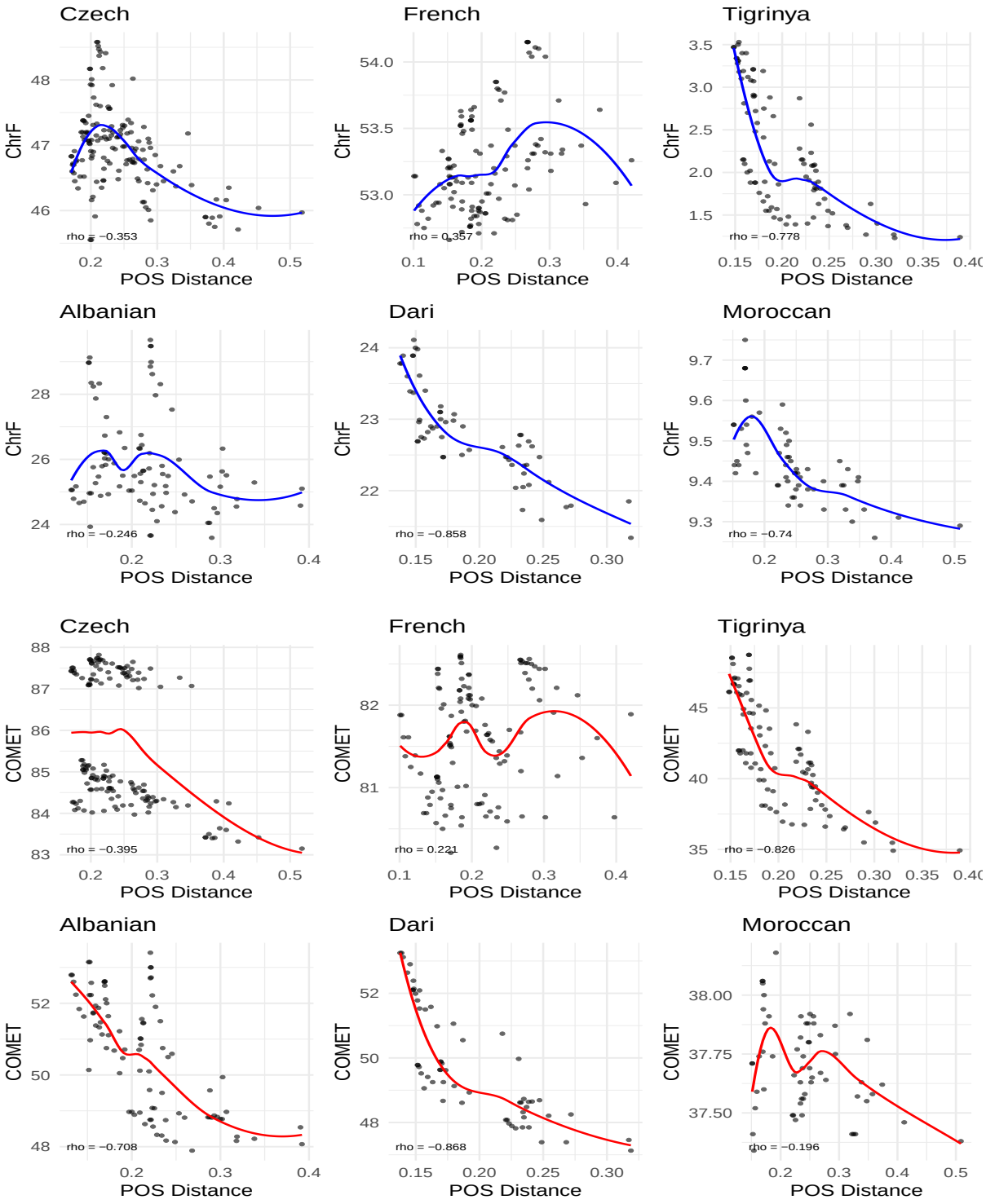


Figure 11: Scatterplots showing the relationship between prompt-level similarity metrics and average translation quality. Each plot corresponds to one of the following prompt-level measurements **POS Distance** (cosine distance between part-of-speech distributions). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

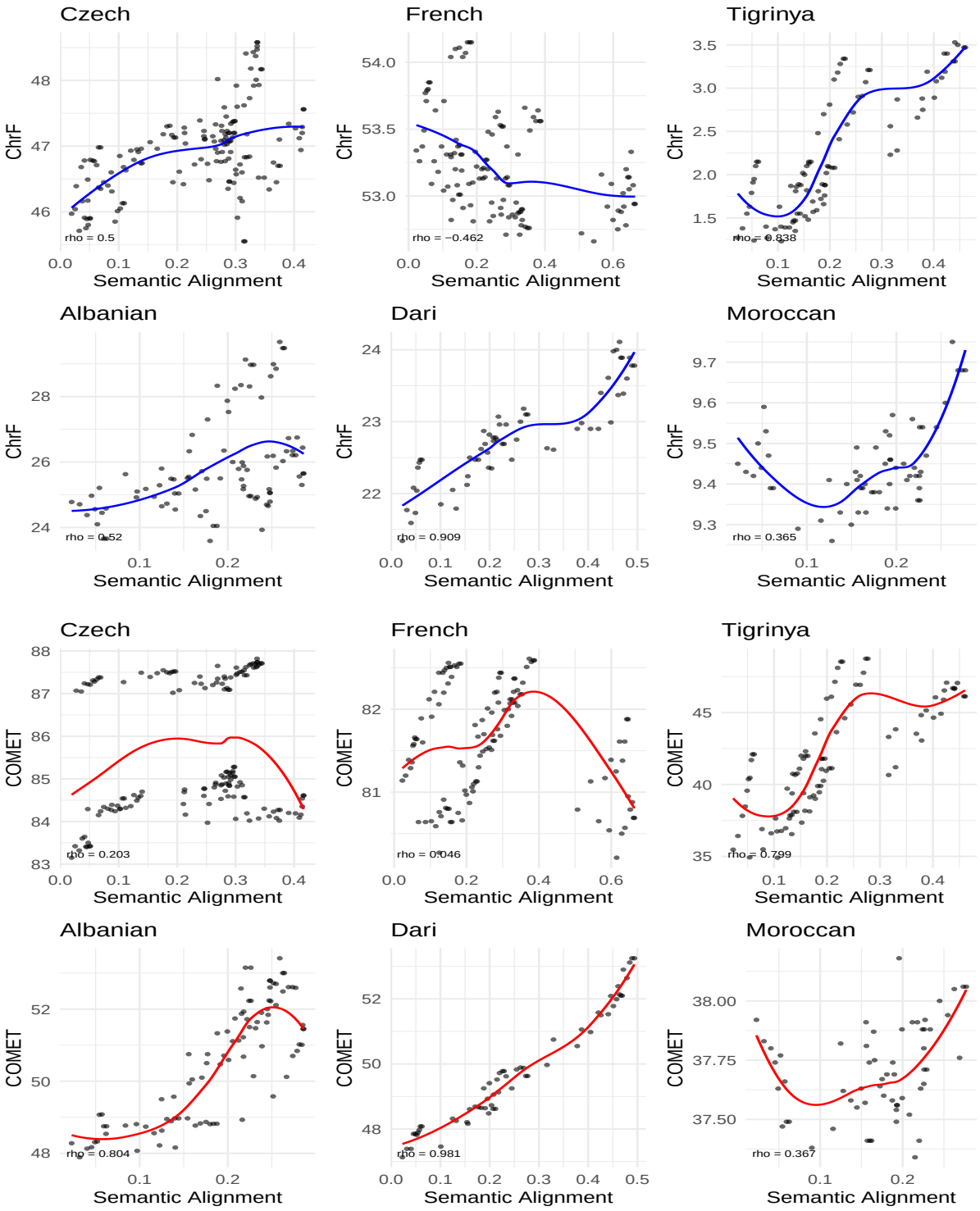


Figure 12: Scatterplots showing the relationship between prompt-level similarity metrics and average translation quality. Each plot corresponds to one of the following prompt-level measurements for **Semantic Alignment** (cosine similarity between sentence embeddings). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.