Multilingual Communication in the Asylum Context: Evaluating LLM-Based Machine Translation with Fuzzy-Match Augmentation and Adaptive NMT across Resource Conditions under Low-Data Constraints

Thomas Moerman, Arda Tezcan and Lieve Macken
Language and Translation Technology Team (LT^3)
Department of Translation, Interpreting and Communication
Ghent University, Belgium
{firstname.lastname}@ugent.be

Abstract

Effective communication in asylum reception settings requires reliable machine translation (MT) across many languages, including low-resource ones. Using data from the MaTIAS project (Machine Translation to Inform Asylum Seekers), we compare retrieval-augmented LLM translation with adaptive Neural MT across 14 target languages with varying resource levels. Working with a very small translation memory of only 358 sentences, we evaluate fuzzy-match (FM) augmentation as an in-context learning strategy for open-source and commercial LLMs and benchmark these against ModernMT with and without domain adaptation. In the LLM setting, FM-based example selection consistently outperforms random selection and zero-shot prompting, with the largest gains for low-resource languages. Adaptive NMT retains an overall advantage, although Gemini Pro approaches its performance and outperforms it on 6 of 14 languages, highlighting a trade-off between translation quality and data sovereignty in privacy-sensitive contexts. These findings show that FM augmentation remains effective under severe data constraints and emphasise the importance of language-specific evaluation in multilingual MT.

1 Introduction

Machine translation (MT) has become an indispensable tool in contexts where linguistic diversity and the need for urgent communication converge, such as in asylum centres, hospitals, and other migrant support services. As migrant populations grow more linguistically diverse – the ten largest groups seeking asylum in Belgium in 2024 alone spoke languages ranging from Arabic and Tigrinya to Pashto and Somali[1] – MT has emerged as a practical means of facilitating communication between migrants and host communities. There is growing evidence that it functions as a primary mediation strategy in these settings: migrants describe relying on tools like Google Translate as their "best friend" for understanding health information and navigating administrative procedures (Valdez and Guerberof-Arenas, 2025), while healthcare professionals report using MT in patient care, albeit largely without institutional guidance or policy frameworks governing its use (Valdez et al., 2025). However, while MT offers the advantage of rapid translation, it has some limitations. Translation quality can be unreliable, and errors are difficult to detect without proficiency in both the source and target languages (Vieira, 2024). Low-resource languages, in particular, suffer from limited data availability, resulting in inaccurate or even incomprehensible translations (Macken et al., 2025).

These challenges closely align with the Human-Centred AI Language Technology (HCAILT) framework introduced by Briva-Iglesias and O'Brien (2026). This framework foregrounds reliability, safety and trustworthiness as foundational principles in the design and deployment of language technologies. One of the framework's two fundamental drivers is the augmentation of information dissemination through rapid, accurate, and context-appropriate communication across lin-

[1] https://www.fedasil.be/nl/downloads/jaarverslag-2024

guistic boundaries – a goal that is directly at stake in migration settings. To improve reliability, HCAILT specifically advocates implementing retrieval-augmented generation (RAG), a method in which an AI system draws on trusted external sources to improve its outputs. However, the authors also acknowledge a central, unresolved tension: ensuring consistent reliability across languages and domains remains challenging due to the scarcity of data for less commonly spoken languages, leading to disparities in the quality of AI-based communication.

It is precisely this tension – ensuring consistent reliability across languages, particularly for low-resource languages – that the present study seeks to address. Focusing on the domain of asylum reception communication, we draw on data from MATIAS, the Machine Translation to Inform Asylum Seekers project (Macken et al., 2025), to evaluate two distinct paradigms for integrating domain-specific knowledge into the translation workflow. The first is adaptive neural MT (adaptive NMT), a set of techniques that allow NMT systems to adapt to domain-specific data, often through fine-tuning or online learning. The second is LLM-based MT with fuzzy-match (FM) augmentation, a form of retrieval-augmented translation in which relevant matches from a translation memory (TM) are retrieved and used to guide the large language model's output. By comparing both approaches across 14 target languages of varying levels of resourcedness, including several low-resource languages, we aim to assess how well each paradigm meets the reliability goals outlined in HCAILT. We also benchmark both approaches against baseline model performance without domain adaptation to quantify the gains that domain-specific integration can provide.

The contributions of this study are as follows:

- We demonstrate that augmenting LLM prompts with FMs consistently improves translation performance across low-, mid-, and high-resource target languages under extremely low-data conditions, compared with zero-shot prompting and randomly selected in-context examples.

- We find that the general-purpose commercial LLM Gemini Pro, through FM augmentation, substantially outperforms smaller, translation-specific open-source models and approaches the performance of adaptive NMT. This highlights the trade-off between translation quality and data sovereignty, which is particularly relevant in privacy-sensitive deployment contexts.

- We show that the relative advantage of FM-augmented LLMs over adaptive NMT is strongly modulated by language resourcedness: while LLMs approach NMT quality for high-resource languages, NMT retains a substantial advantage for most low-resource languages – with notable exceptions such as Tigrinya.

2 Related research

2.1 Language resourcedness

Neural approaches to MT, whether traditional NMT systems or more recent LLM-based applications, are inherently data-driven. Their performance depends on the volume and quality of the available training data for a given language pair. Yet progress has largely benefited only a handful of high-resource languages, leaving many others behind. This imbalance is starkly illustrated by Meta's No Language Left Behind (NLLB) project (NLLB Team et al., 2022), which aims to provide high-quality MT for 200 languages. NLLB defines 'low-resource' using the criterion of having fewer than one million publicly available, deduplicated bitext samples. Under this criterion, 150 of the 200 NLLB target languages qualify as low-resource.

However, the binary classification of languages as either 'high-resource' or 'low-resource' has faced growing criticism for being overly simplistic. Joshi et al. (2020) therefore propose a more nuanced system, categorising over 2,500 languages into six tiers (0–5) based on two criteria: the availability of unlabelled data (Wikipedia page counts) and labelled data (resources from the LDC catalogue and ELRA Map). Furthermore, Keet and Khumalo (2023) argue that corpus size alone is insufficient to capture a language's full resourcedness. Their five-level matrix – ranging from 'Very Low-Resource' to 'Very High-Resource' – considers eleven contextual factors, such as grammar documentation, educational support, funding, policy environments, and the availability of human language technology tools. In their scheme, only English qualifies as 'Very High-Resource'.

Resourcedness is not only an MT training issue. It also affects the reliability (and availability) of automatic evaluation metrics, such as COMET (Rei et al., 2020), which are widely used to assess translation quality. As these metrics are based on neural models themselves, they may produce less reliable scores for lower-resource languages.

2.2 Integrating domain-specific knowledge into the translation workflow

Both NMT systems and LLMs can leverage domain-specific knowledge through fuzzy-matches (FMs), that is, translations similar to a given input retrieved from a translation memory (TM) or parallel corpus, but they do so in different ways.

In the NMT paradigm, several approaches have been proposed to integrate FMs. Early work on fuzzy-match repair (FMR) took a direct approach, algorithmically patching mismatched subsegments in retrieved FM targets using an external MT system (Ortega et al., 2016). Subsequent methods range from architectural modifications such as additional attention layers (Cao and Xiong, 2018; He et al., 2021), memory components (Feng et al., 2017), or FM-editing architectures (Bouthors et al., 2023), to data augmentation techniques that leave the model architecture unchanged. One example of the latter is Neural Fuzzy Repair (NFR) (Bulté and Tezcan, 2019), in which the source sentence is concatenated with the target side of its most similar FM at both training and inference time. This approach has proven effective in domain-specific scenarios with sufficient parallel data for both training and FM retrieval (with a minimum of approximately 300K sentence pairs), and subsequent work has confirmed its usefulness across additional language pairs and domains (Xu et al., 2020; Tezcan et al., 2021). A consistent finding is that FM augmentation quality depends on the similarity of retrieved FMs to the input and on the size of the available data pool (Bulté and Tezcan, 2019; Xu et al., 2020; Tezcan et al., 2021).

A complementary approach to domain adaptation in NMT is instance-based adaptation, in which the model itself is dynamically adjusted at inference time using retrieved similar sentences. Farajian et al. (2017) proposed an unsupervised method that, for each input sentence, retrieves similar sentence pairs from the training data and uses them to locally fine-tune the NMT model on-the-fly. This approach was integrated into the open-source ModernMT system (Bertoldi et al., 2018), which combines instance-based adaptation with a dynamic memory that stores user-provided TMs. Unlike NFR, which modifies the model input via source augmentation, ModernMT[2] adapts the model parameters for each translation request, making it particularly well-suited to multi-domain scenarios in which a single system must handle diverse content. It should be noted that the publicly available open-source version has not been updated since 2021, and the system has since been acquired by Translated. The exact relationship between the open-source release and the current proprietary version is unclear. ModernMT serves as the adaptive NMT baseline in the present study, as it is deployed in the MaTIAS prototype (Macken et al., 2025).

LLMs, by contrast, can integrate FMs without retraining. In an in-context learning (ICL) setting, retrieved FMs are provided as few-shot examples directly in the prompt, allowing the model to adapt its output without parameter updates. Moslem et al. (2023a) showed that this approach significantly improves translation quality using a retrieval pool of 3,070 segments, with performance depending more on the semantic similarity of selected examples than on their quantity. The choice of example selection strategy is itself an active area of investigation: while similarity-based retrieval is the most common approach, randomly selected examples have also been shown to improve translation quality over zero-shot prompting (Alves et al., 2023). While LLMs have shown strong general-domain MT performance (Kocmi et al., 2023; Kocmi et al., 2024), their effectiveness in specialised domains remains less conclusive, with NMT systems maintaining competitive results in areas such as biomedical and patent translation (Higashiyama, 2024; Neves et al., 2024; Wassie et al., 2024).. Recent work has also begun to explore the scaling properties of ICL for MT: Salim et al. (2026) scale in-context demonstrations up to 1M tokens for low-resource language translation, finding that gains saturate early and can degrade near the maximum context window, with scaling behaviour strongly dependent on corpus type.

In summary, the approaches discussed above

[2] https://github.com/modernmt/modernmt

thus differ in key aspects. Adaptive NMT through FM augmentation requires training a dedicated model on domain-specific data and is tightly coupled to the availability of large parallel corpora. Instance-based adaptation, as implemented in ModernMT, is more flexible because it operates on-the-fly without retraining, but it still requires a pool of parallel data for retrieval and local fine-tuning, and its adaptation is transient rather than cumulative. LLM-based FM augmentation through ICL, by contrast, requires no task-specific training or fine-tuning, making it immediately deployable even with limited parallel data, albeit at the cost of relying entirely on the base model's inherent capabilities. Despite these complementary strengths, direct comparative evaluations between different paradigms remain scarce. Moerman et al. (2025) compared FM-augmented NMT with LLMs prompted with FMs for scientific literature translation (English–French) with TMs ranging from 100K to 44M sentence pairs, finding that the best NMT configuration surpassed smaller LLMs on lexical overlap metrics while a larger LLM achieved the highest COMET scores. However, a comprehensive comparison across multiple language pairs and resource conditions, including low-resource languages, has not yet been conducted. The present study aims to help fill this gap.

3 Dataset

The dataset used in this study comprises 200 English messages, each of which is professionally translated into 14 target languages (Table 2). These messages originate from the MaTIAS project and are intended for quick, clear communication, focusing on logistical and administrative updates in asylum reception centres. Examples include reminders like "It is not allowed to leave bicycles in the corridor!" or friendly follow-ups such as "Last week, you joined us on a trip to the museum. Check out the photos of the activity here!". Each message can contain multiple sentences, resulting in a total of 781 sentences across the 200 messages. The dataset was divided into two equal subsets[3]: 100 messages (358 sentences, totalling 3,193 English words) were used to create a context-specific translation memory (TM) for FM retrieval, while the remaining 100 messages (423 sentences, 3,518 English words) were allocated for testing translation performance (Table 1). This 50/50 split was motivated by the dataset's small total size: reserving a larger portion for the TM would have left too few messages for reliable evaluation, and vice versa. While prior work on FM augmentation for both NMT and LLMs has typically relied on relatively large retrieval pools, often containing thousands or even hundreds of thousands of sentence pairs (Bulté and Tezcan, 2019; Moslem et al., 2023a), the effect of FM-based ICL with extremely small TMs has not been explored. Our setting, with a TM of only 358 sentences, thus represents a novel and practically relevant scenario for investigating retrieval-augmented LLM translation under severe data constraints. It should be noted that this small dataset size is not an artificial experimental choice but reflects the reality of the MaTIAS project, where the total volume of professionally translated material was inherently limited by the project's scope and resources.

Statistic	Train (TM)	Test
Messages	100	100
Sentences	358	423
Words	3,193	3,518

Table 1: Source language (English) dataset statistics.

Table 2 provides an overview of the 14 target languages included in this study. To characterise language resourcedness, we draw on two established classifications. The Joshi class (0–5) follows the taxonomy proposed by Joshi et al. (2020), which categorises languages according to the availability of digital resources and NLP tools. The NLLB column indicates the resource level assigned by Meta's No Language Left Behind (NLLB) project (NLLB Team et al., 2022). Based on these two classifications, we assign each language to a Tier used in this study, grouping languages into broader resource-level categories by considering both the Joshi and NLLB labels.

4 Method and experimental setup

4.1 Models

Following previous work, we compare two paradigms for domain-specific MT. As the **adaptive NMT** baseline, we use ModernMT[4], a context-aware, incremental NMT system based on the Fairseq Transformer architecture that can leverage a domain-specific TM at inference

[3]Even vs. odd numbered messages.

[4]`https://www.modernmt.com/`

Language	ISO	Script	Joshi	NLLB	Tier
Arabic	ar	Arabic	5	High	High
German	de	Latin	5	High	High
Spanish	es	Latin	5	High	High
Farsi	fa	Arabic	4	High	High
Portuguese	pt	Latin	5	High	High
Romanian	ro	Latin	5	High	High
Russian	ru	Cyrillic	5	High	High
Turkish	tr	Latin	5	High	High
Armenian	hy	Armenian	2	Low	Mid
Albanian	sq	Latin	1	High	Mid
Georgian	ka	Georgian	2	Low	Mid
Pashto	ps	Arabic	1	Low	Low
Somali	so	Latin	1	Low	Low
Tigrinya	ti	Ge'ez	1	Low	Low

Table 2: Overview of the 14 target languages, including their Joshi class, NLLB resource level, and the Tier grouping used in this study.

time. This system was deployed in the MaTIAS project (Macken et al., 2025) and serves as the production baseline. We evaluate ModernMT both with and without the context-specific TM described in Section 3. Like API-based LLMs, ModernMT offers deployment without requiring local model maintenance, GPU infrastructure, or the implementation of retrieval pipelines or prompting strategies.

For **LLM-based MT with FM augmentation**, we evaluate five models in a few-shot in-context learning (ICL) setup, where translation examples retrieved from the TM are provided as part of the input prompt. Four are open-source, translation-specific models: three variants of Google's TranslateGemma (Finkelstein et al., 2026) (4b, 12b, and 27b parameters; hereafter **TG-4b**, **TG-12b**, and **TG-27b**) and Unbabel's TowerPlus-9b (Rei et al., 2025) (**TP-9b**). The fifth is a closed-source, general-purpose model from Google's Gemini 2.5 family (Gemini Team, Google, 2025) (Pro; hereafter **GP**), which scored among the top systems in recent WMT shared tasks (Kocmi et al., 2025b; Kocmi et al., 2025a). We refer to ModernMT with TM as **MMT$^+$** and without TM as **MMT**.

4.2 FM retrieval

For FM retrieval, we follow the established approach described in the literature (Moslem et al., 2023a; Tezcan et al., 2024): all source sentences in the TM are encoded using multilingual sentence embeddings from the Multilingual-MiniLM-L12-H384 model (Wang et al., 2021) and indexed with FAISS (Johnson et al., 2021) using an IVFFlat index with Euclidean distance. At inference time, the k nearest TM entries for each source sentence are retrieved and provided as source–target examples in the prompt, ranked by increasing distance (i.e., most similar first; see Appendix A and Table 3). Exact matches (identical lowercased strings or near-zero distance) are filtered out to avoid augmenting a sentence with itself. Since retrieval operates on the English source side, the same examples are retrieved for all 14 target languages. As a comparison condition, we also evaluate random example selection, in which examples are drawn randomly from the TM rather than based on similarity. Both strategies are compared against a zero-shot baseline (no in-context examples). All stages of the pipeline, including translation, retrieval, and evaluation, are performed at the sentence level.

Table 3 illustrates the difference between fuzzy-match (FM) retrieval and random example selection for the source sentence "You have an appointment at the medical unit on 2 October at 10:00." FM retrieval selects semantically similar sentences – sharing vocabulary about appointments, times, and medical services – while random selection draws unrelated examples from the TM. For each retrieved example, both the source sentence and its target translation are provided to the model as a source–target pair in the prompt (see Appendix A for the full prompt template).

Source: *You have an appointment at the medical unit on 2 October at 10:00.*	
FM retrieval (ranked by L2 distance, lower = more similar)	
0.23	You have a **dental appointment** at **10:00** on **2 October**.
0.27	You have an **appointment** with your social worker on Monday **5 October at 14:00**.
0.55	The nurse will see you on the same day between **10:15 and 12:00**.
Random selection	
–	You must deliver your contract to the Talent Twister department at the end of the week.
–	From 14:00 to 15:00 on the fourth floor.
–	Children's Choir On-Track

Table 3: Examples of FM retrieval vs random selection for a single source sentence. L2 distance scores are shown for FM retrieval (lower values indicate greater similarity). Bold text highlights approximate overlap with the source.

4.3 Evaluation metrics

Translation quality is assessed using three lexical metrics, namely BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and TER (Snover et al.,

2006), and the neural metric COMET, using the `wmt22-comet-da` model (Rei et al., 2020). We report ChrF as the primary lexical metric in the main text, motivated by two considerations: first, ChrF has been adopted as the primary metric in recent WMT shared tasks on terminology translation (Semenov et al., 2025) and automated evaluation (Lavie et al., 2025); second, earlier evaluation (Macken et al., 2025) of this dataset showed that word-based metrics can be unreliable for scripts such as Armenian and Ge'ez (Tigrinya). BLEU and COMET scores are provided in the appendix for all applicable experiments. COMET scores should be interpreted with caution for lower-resource languages such as Pashto, Somali, Georgian, and Tigrinya, where the underlying model has limited training data.

To examine the role of language resourcedness, we group the 14 target languages into three tiers: high-resource, mid-resource, and low-resource, following the categorisation introduced in Table 2.

Statistical significance is assessed using paired bootstrap resampling (Koehn, 2004) with $n = 1{,}000$ resamples drawn with replacement from the test set, comparing the best-performing system against the second-best for each target language. Significance of the best-performing configuration over the second-best is denoted by $^{-}$, *, †, and ‡, representing $p \geq 0.05$, $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

5 Results

5.1 Impact of the number of in-context examples on translation quality

We first examine how the number of FM-based in-context examples (shots) affects translation quality across the four open-source models. Figure 1 shows ChrF scores averaged over all 14 target languages for shot counts ranging from 0 to 20, using FM retrieval.

All four tested models show substantial improvements as the number of FM examples increases, with the steepest gains occurring between 0 and 5 examples. TranslateGemma-27b improves from 55.77 (zero-shot) to 61.34 ChrF (5 examples) and reaches its peak of 62.32 at 12 examples before declining slightly at higher counts. TranslateGemma-12b follows a similar trajectory, peaking at 61.17 (18 examples), while TranslateGemma-4b plateaus more broadly between 12 and 20 examples (around 55.6–55.8 ChrF). The larger models reach their optimum earlier than smaller models and then decline, suggesting that overly long prompts may introduce noise or exceed optimal context utilisation. TowerPlus-9b shows a similar scaling pattern but with consistently lower absolute scores, reflecting its limited language coverage (only 5 of our 14 target languages are supported). Similar patterns hold for BLEU and COMET (Appendix B), where larger models consistently achieve the highest scores, and the same saturation behaviour is observed.

Across all four models, performance plateaus between 12 and 18 examples. We fix the number of in-context examples at **15** for all subsequent experiments, as this captures most of the gains while avoiding the degradation observed at higher counts. This is broadly consistent with prior work: Moslem et al. (2023a) used up to 10 FM-based examples, while Alves et al. (2023) used 5 randomly selected examples in a fine-tuning context. More recently, Salim et al. (2026) scaled ICL demonstrations up to 1M tokens for low-resource MT, finding that performance saturates around 2^{14}–2^{16} tokens and can degrade near the maximum context window. While their setting differs from ours (long-context models, larger demonstration pools), the shared finding that more context does not always yield better results reinforces our choice of a moderate number of high-quality examples.

5.2 Example selection strategies: fuzzy-match retrieval versus random sampling

To isolate the effect of the example selection strategy, we compare examples retrieved using FMs with those selected at random, as well as with a zero-shot baseline that uses TranslateGemma-27b in a fixed 15-example setting. We use the 27b variant as this model yields the highest translation performance in the previous experiment (Section 5.1). Figure 2 shows ChrF scores per target language for all three conditions, with languages grouped by resource tier.

The zero-shot baseline (average ChrF 55.77) establishes the model's intrinsic translation capability. Both example selection strategies substantially improve over this baseline: FM retrieval yields an average improvement of +6.29 ChrF, while random selection yields +3.34 ChrF. FM retrieval outperforms random selection for 13 of 14 target languages, and this advantage is statistically significant for 11 of 14 languages ($p < 0.01$); the

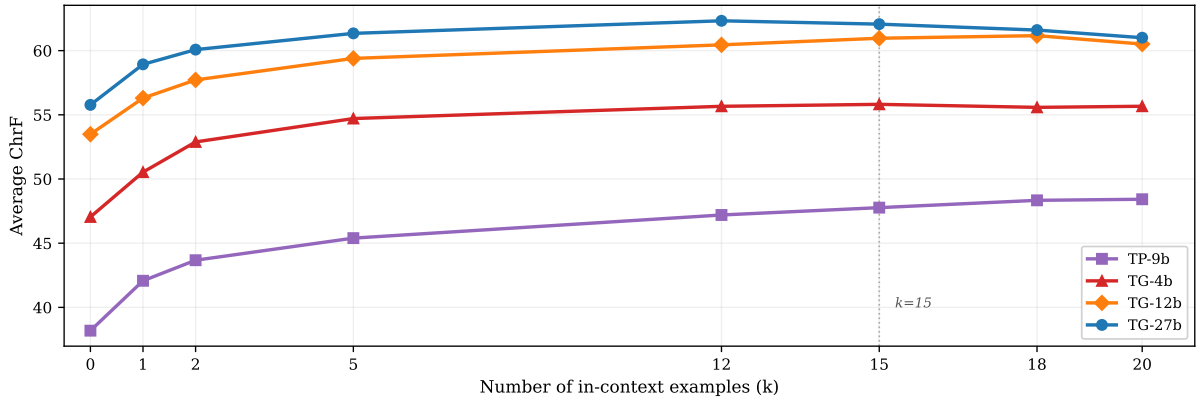

Figure 1: Average ChrF as a function of shot count for the four open-source models – TranslateGemma-27b (TG-27b), TranslateGemma-12b (TG-12b), TranslateGemma-4b (TG-4b), and TowerPlus-9b (TP-9b) – with fuzzy-match retrieval. The dashed line marks the fixed shot count of 15 used in subsequent experiments.

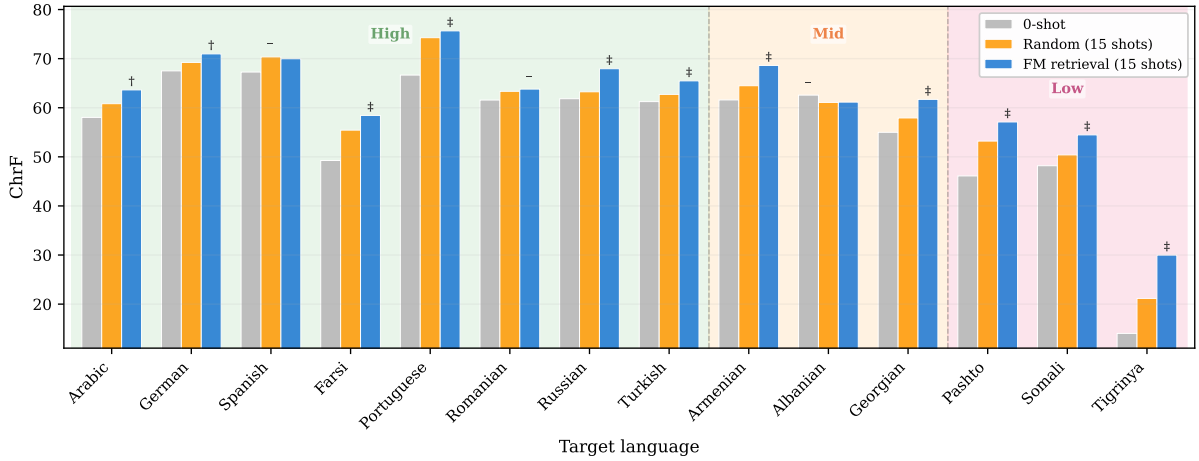

Figure 2: ChrF scores per target language for zero-shot, random selection, and fuzzy-match retrieval (TranslateGemma-27b, 15 examples). Languages ordered by resource tier. Significance markers ($^-$/∗/†/‡) above the best bar denote paired bootstrap test results comparing the best strategy against the second-best per language, representing $p \geq 0.05$, $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

three exceptions – Spanish, Romanian, and Albanian – all have absolute gaps of less than 1.5 ChrF points. Spanish is the only language in which random selection marginally outperforms FM retrieval (-0.37 ChrF); this difference is not statistically significant. The advantage of FM retrieval over random selection is also limited for several other high-resource languages, including Romanian (+0.45) and Portuguese (+1.39), suggesting that models with strong intrinsic capabilities for these languages are less sensitive to example quality. By contrast, the largest FM advantages are observed for Russian (+4.70) and Farsi (+2.99), languages for which the model may have less domain-specific knowledge.

The benefit of FM retrieval varies substantially across resource tiers. For the **high-resource** group, FM retrieval improves over random selection by an average of +2.06 ChrF (66.98 vs 64.93). For the **mid-resource** group (Armenian, Albanian, Georgian), the average advantage is +2.66 ChrF (63.82 vs 61.16), though the overall gain from any in-context examples over the zero-shot baseline is more modest (average +4.10 ChrF for FM over zero-shot), suggesting these languages benefit less from ICL in general but still distinguish between example quality. The **low-resource** group (Pashto, Somali, Tigrinya) shows the most striking pattern: FM retrieval outperforms random by an average of +5.60 ChrF (47.19 vs 41.59), more than double the advantage observed for high-resource languages. These languages also show the largest absolute gains from any form of augmentation (average +11.07 ChrF for FM over zero-shot), confirming that both example quantity and quality matter most where the model's intrinsic capabil-

ities are weakest. Tigrinya exemplifies this most clearly: FM augmentation more than doubles the ChrF score over the zero-shot baseline (29.99 vs 14.04) and outperforms random selection by +8.83 ChrF.

These results are particularly remarkable given the small size of the retrieval pool: the TM from which examples are selected contains only 358 sentences. Prior work on FM augmentation for NMT has typically relied on retrieval pools of at least 300K sentence pairs (Bulté and Tezcan, 2019; Xu et al., 2020), while LLM-based retrieval studies have used pools ranging from 3,070 segments (Moslem et al., 2023a) and 50,000 segments (Moslem et al., 2023b) to 100K–44M sentence pairs (Moerman et al., 2025). Our results demonstrate that, even with a drastically smaller dataset, similarity-based example selection consistently identifies examples more useful than randomly selected ones, confirming that FM augmentation is effective even in severely resource-constrained scenarios. Similar patterns are observed for BLEU and COMET (Appendix B).

5.3 LLM-based translation vs adaptive NMT across languages

In this section, we compare the best open-source LLM (TranslateGemma-27b, 15 examples with FM retrieval) against both commercial systems: ModernMT (with and without TM) and Gemini Pro (15 examples with FM retrieval). Table 4 presents ChrF scores per target language for all four systems, with languages grouped by resource tier. Note that FM augmentation through ICL applies only to the LLMs; ModernMT integrates domain knowledge through its TM-based instance adaptation mechanism.

Overall comparison. ModernMT with TM (MMT$^+$) achieves the highest overall ChrF (66.60) and the best scores for 8 of 14 languages. The TM provides a consistent boost: MMT$^+$ outperforms MMT by +3.18 ChrF on average, with gains for 13 of 14 languages (the exception being Georgian, where MMT without TM outperforms MMT$^+$ by +3.87). Among the LLMs, Gemini Pro (65.38) substantially outperforms TranslateGemma-27b (62.06) and trails MMT$^+$ by only -1.22 ChrF, matching or outperforming it on 6 of 14 languages. TranslateGemma-27b trails MMT$^+$ by -4.54 ChrF, outperforming it on only 3 languages (Farsi, Russian, Turkish). However, 6 of 14 per-language ChrF differences between the best and second-best system are not statistically significant (German, Spanish, Russian, Turkish, Armenian, Albanian), all with gaps ≤ 1.66 ChrF. Languages where MMT$^+$ remains clearly superior to both LLMs include Georgian, Pashto, Portuguese, and Arabic. Notably, absolute translation quality does not always correlate with resource tier: Georgian achieves 79.87 ChrF with MMT$^+$ (higher than all high-resource languages except Portuguese), while the high-resource language Farsi scores only 54.01. This divergence possibly reflects both script properties and morphological typology. Georgian's phonemic Mkhedruli script has a small, fixed character inventory (33 letters, no contextual forms), which favours clean character-level alignments in ChrF. Farsi's Arabic-derived script, by contrast, features contextual letter forms and optional diacritics that introduce variability in encoding. Additionally, Georgian's agglutinative morphology may produce more predictable character n-gram patterns under NMT, while Farsi's more fusional structure interacts differently with subword tokenisation in LLMs.

	Target	MMT$^+$	MMT	TG-27b	GP
High	Arabic	**66.76**‡	63.78	63.63	63.76
	German	73.67	69.26	70.97	**73.85**$^-$
	Spanish	72.64	69.77	69.98	**74.30**$^-$
	Farsi	54.01	51.36	**58.43**‡	55.03
	Portuguese	**80.97**‡	75.48	75.65	76.54
	Romanian	**67.97**†	63.10	63.80	67.22
	Russian	67.62	65.25	**67.95**$^-$	67.01
	Turkish	65.34	62.20	65.47	**65.60**$^-$
	Avg	*68.62*	*65.03*	*66.98*	*67.91*
Mid	Armenian	**74.26**$^-$	68.58	68.61	73.42
	Albanian	70.01	65.46	61.15	**70.05**$^-$
	Georgian	79.87	**83.74**‡	61.70	65.69
	Avg	*74.72*	*72.59*	*63.82*	*69.72*
Low	Pashto	**66.82**‡	64.86	57.10	62.53
	Somali	**58.19**†	55.57	54.47	57.34
	Tigrinya	34.29	29.48	29.99	**43.00**†
	Avg	*53.10*	*49.97*	*47.19*	*54.29*
	Avg All	**66.60**	**63.42**	**62.06**	**65.38**

Table 4: ChrF scores per target language: ModernMT with TM (MMT$^+$), ModernMT without TM (MMT), TranslateGemma-27b (TG-27b), and Gemini Pro (GP). Both LLMs use 15 in-context examples selected through FM retrieval. Languages are grouped by resource tier. Best scores are in bold; significance markers ($^-$/$*$/†/‡) denote paired bootstrap test results comparing the best system against the second-best per language.

High-resource languages. For the high-resource group, MMT$^+$ leads with an average of 68.62 ChrF. TranslateGemma-27b is competitive, trailing by -1.64 ChrF on average, while Gemini Pro narrows the gap further to -0.71. Farsi is a clear LLM winner, while Arabic and Portuguese favour MMT$^+$. The TM benefit for MMT is substantial for this group, averaging +3.60 ChrF.

Mid-resource languages. MMT$^+$ leads with an average of 74.72 ChrF. TranslateGemma-27b trails by -10.90 ChrF, with large deficits for Georgian (-18.18) and Albanian (-8.86); the latter likely reflects limited LLM pretraining data (Joshi class 1) despite its Latin script, which is generally better represented in LLM training data, yet script familiarity alone does not compensate for insufficient language-specific data. Gemini Pro closes the gap, trailing by -5.00 ChrF. Georgian is particularly striking: MMT without TM achieves 83.74 ChrF – the highest score for any language-system combination – while both LLMs fall far below.

Low-resource languages. The low-resource group (Pashto, Somali, Tigrinya) reveals the most complex patterns. MMT$^+$ retains a clear ChrF advantage for Pashto (-9.72 for TranslateGemma-27b, -4.28 for Gemini Pro). For Somali, the gap is smaller but still favours MMT$^+$. However, for Tigrinya – the lowest-resource language in our set – Gemini Pro reverses the pattern entirely, outperforming MMT$^+$ by +8.70 ChrF, while TranslateGemma-27b trails by -4.31. Overall, Gemini Pro slightly outperforms MMT$^+$ for this group on average (54.29 vs 53.10). These results demonstrate that 'low-resource' is not a monolithic category: the LLM–NMT comparison depends on factors including the language's representation in LLM pretraining data, its script, and the specific domain.

Metric agreement. Overall, ChrF and BLEU are strongly correlated across languages (Spearman $\rho = 0.90$), while the ChrF–COMET correlation is moderate ($\rho = 0.77$) and drops to near zero for the low-resource group, indicating that these metrics capture different quality dimensions for underrepresented languages. One notable disagreement occurs for Georgian, where ChrF strongly favours ModernMT over TranslateGemma-27b ($+18.18$) while COMET slightly favours the LLM. For Pashto and Somali, all three metrics agree in their system rankings. These patterns underscore the importance of reporting multiple metrics, particularly for lower-resource languages where neural evaluation metrics may be less reliable (see Appendix B for full BLEU and COMET results).

6 Discussion and conclusion

Our findings can be situated within the broader HCAILT framework (Briva-Iglesias and O'Brien, 2026), which identifies reliability, safety, and trustworthiness as foundational requirements for language technologies deployed in high-stakes communication settings. The results of our experiments consistently support the central hypothesis of this study: integrating domain-specific knowledge retrieved from TMs improves translation quality, regardless of whether the underlying system is an adaptive NMT model or an LLM. This finding aligns with the broader argument put forward by HCAILT that RAG approaches represent a promising path toward more reliable AI-assisted communication.

When comparing the two paradigms examined in this study, augmenting LLM prompts with FMs retrieved from a TM consistently outperforms both zero-shot prompting and randomly selected in-context examples across low-, mid-, and high-resource target languages. This effect is particularly noteworthy, as FM augmentation yields consistent improvements across all languages, even under extremely low-data conditions. At the same time, adaptive NMT remains a competitive approach, outperforming several LLM-based configurations.

Despite these encouraging results, the case of Tigrinya highlights the limitations of current approaches to extremely low-resource languages. Although the Gemini Pro model produced promising improvements for Tigrinya compared to other evaluated systems, translation quality remains considerably lower than for other target languages in our study. Crucially, a multilingual quality evaluation conducted within the MaTIAS project (Macken et al., 2025), which assessed ModernMT translations across the project's target languages by bilingual human experts, found that the quality of Tigrinya MT was deemed too low for end-user use. It remains to be seen whether the obtained quality improvements will change this in real-world deployment.

A further finding that complicates the standard

framing of the low-resource problem is Farsi's underperformance. Despite being categorised as a high-resource language, Farsi achieved the second-lowest translation quality scores in our experiments. More broadly, although general trends based on language resourcedness were observable, the LLM–NMT comparison revealed that factors such as representation in training data, script, and the interplay between language and domain all contribute to shaping translation quality in ways that resourcedness alone cannot capture.

While this study has focused primarily on reliability, the HCAILT framework identifies safety as an equally fundamental requirement. In the context of AI-assisted communication, safety encompasses strict privacy and data security protocols – considerations that are of particular importance in sensitive environments such as asylum centres and hospitals. Open-source models offer a distinct practical advantage in this regard, as they enable developers to retain full control over data handling and avoid routing sensitive information through commercial APIs. However, our results show that the translation-specific open-source models evaluated here performed worse than the commercial systems, creating a tension between safety and reliability that current tools have yet to resolve.

Taken together, these findings underscore both the promise and the remaining challenges of deploying AI-assisted translation in low-resource settings. Domain-specific knowledge integration – whether through adaptive NMT or retrieval-augmented LLMs – represents a meaningful step toward the reliability that frameworks such as HCAILT foreground. Yet the case of Tigrinya serves as a reminder that technical improvements might not automatically translate into usable, equitable tools for all language communities. This concern is echoed by Mager et al. (2023), who, through interviews with Indigenous language community members, found that low-quality MT is perceived as potentially harmful and that community consultation is essential when deploying MT for underrepresented languages. Beyond reliability, the unresolved tension between safety and performance underscores a broader need for the field to develop open-source models that meet the quality bar set by commercial systems while safeguarding data security. Addressing these challenges will require not only continued technical innovation but also sustained investment in data collection for underrepresented languages.

7 Limitations

Several limitations should be acknowledged. First, the evaluation relies exclusively on automatic metrics. While we report three complementary metrics (ChrF, BLEU, COMET), none fully captures translation quality as perceived by end users, particularly for low-resource languages where metric reliability is itself uncertain. Human evaluation would be needed to validate these findings. Conducting such an evaluation presents particular challenges in this setting: it requires recruiting qualified evaluators for all 14 target languages, including low-resource languages such as Pashto, Somali, and Tigrinya, for which finding domain-knowledgeable bilingual evaluators is exceptionally difficult. A prior human evaluation within the MaTIAS project (Macken et al., 2025), which assessed ModernMT translations, confirmed both the value and the practical constraints of user-centred quality assessment in this multilingual context.

Second, the dataset comprises 200 messages (781 sentences) from a single domain (asylum reception communication), which limits generalisability to other domains or text types. The small TM of 358 sentences, while representative of real-world low-data scenarios, constrains FM retrieval quality due to limited coverage.

Third, our comparison is limited to a specific set of models evaluated at a single point in time. LLM capabilities evolve rapidly, and newer model versions may narrow or widen the gaps observed here. Similarly, the commercial models (ModernMT and Gemini Pro) were accessed via API, meaning results may not be exactly reproducible if the underlying models are updated.

Fourth, COMET scores should be interpreted with particular caution for the low-resource languages in our study. As a neural metric trained predominantly on higher-resource language data, COMET may not reliably reflect translation quality for languages such as Pashto, Somali, and Tigrinya, as evidenced by the near-zero ChrF–COMET correlation observed for this group.

Finally, our study does not address computational cost, latency, or deployment complexity in detail. While we note the trade-off between data sovereignty and translation quality, a full cost-benefit analysis – including inference time, API costs, and infrastructure requirements – would be

valuable for practitioners considering deployment in similar settings.

8 Carbon impact statement

We report the computational resources used in this study to enable assessment of its carbon footprint, following the recommendations of Strubell et al. (2019). Inference for the four open-source models (TranslateGemma-4b/12b/27b and TowerPlus-9b) was performed on two NVIDIA A100 80GB GPUs through a vLLM instance[5]. The commercial models Gemini Pro and ModernMT were accessed through their respective APIs, for which energy consumption is not directly measurable. No model training or fine-tuning was performed in this study; all experiments involved inference only. Given the relatively small dataset (423 test sentences $\times$ 14 languages $\times$ 15 shots) and the inference-only nature of the experiments, the overall computational footprint is modest compared to studies involving model training.

Acknowledgements: The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), which is funded by Ghent University, FWO and the Flemish Government department EWI.

References

Alves, Duarte M., Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore, December. Association for Computational Linguistics.

Bertoldi, Nicola, Davide Caroselli, and Marcello Federico. 2018. The ModernMT project. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 365, Alicante, Spain, May. European Association for Machine Translation.

Bouthors, Maxime, Josep Crego, and François Yvon. 2023. Towards example-based NMT with multi-Levenshtein transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1846, Singapore, December. Association for Computational Linguistics.

[5] https://github.com/vllm-project/vllm

Briva-Iglesias, Vicent and Sharon O'Brien. 2026. Human-centered AI language technology (HCAILT): An empathetic design framework for reliable, safe and trustworthy multilingual communication. *International Journal of Human–Computer Interaction*, pages 1–15. Advance online publication.

Bulté, Bram and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy, July. Association for Computational Linguistics.

Cao, Qian and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047, Brussels, Belgium, October-November. Association for Computational Linguistics.

Farajian, M. Amin, 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, September. Association for Computational Linguistics.

Feng, Yang, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark, September. Association for Computational Linguistics.

Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma technical report. *arXiv preprint arXiv:2601.09012*.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

He, Qiuxiang, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, Online, August. Association for Computational Linguistics.

Higashiyama, Shohei. 2024. Results of the WAT/WMT 2024 shared task on patent translation. In *Proceedings of the Ninth Conference on Machine*

Translation, pages 118–123, Miami, Florida, USA, November. Association for Computational Linguistics.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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

Keet, C. Maria and Langa Khumalo. 2023. Contextualising levels of language resourcedness that affect NLP tasks. *arXiv preprint arXiv:2309.17035*.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

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

Kocmi, Tom, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025a. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China, November. Association for Computational Linguistics.

Kocmi, Tom, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China, November. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China, November. Association for Computational Linguistics.

Macken, Lieve, Margot Fonteyne, Arda Tezcan, Ella van Hest, Katrijn Maryns, and July De Wilde. 2025. Machine translation in asylum reception centres: system selection and multilingual quality evaluation. *Revista Tradumàtica: translation technologies*, (23):326–349.

Mager, Manuel, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada, July. Association for Computational Linguistics.

Moerman, Thomas, Tom Vanallemeersch, Sara Szoc, and Arda Tezcan. 2025. Tailoring machine translation for scientific literature through topic filtering and fuzzy match augmentation. In Tsunakawa, Takashi, Katsuhito Sudoh, and Isao Goto, editors, *Proceedings of the Eleventh Workshop on Patent and Scientific Literature Translation (PSLT 2025)*, pages 13–26, Geneva, Switzerland, June. European Association for Machine Translation.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. 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, June. European Association for Machine Translation.

Moslem, Yasmin, Rejwanul Haque, and Andy Way. 2023b. Fine-tuning large language models for adaptive machine translation. *arXiv preprint arXiv:2312.12740*.

Neves, Mariana, 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, November. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Ortega, John, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2016. Fuzzy-match repair using black-box machine translation systems: what can be expected? In Green, Spence and Lane Schwartz, editors, *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 27–39, Austin, TX, USA, October. The Association for Machine Translation in the Americas.

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

Popović, Maja. 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, September. Association for Computational Linguistics.

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

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual LLMs. *arXiv preprint arXiv:2506.17080*.

Salim, Luis Frentzen, Esteban Carlin, Alexandre Morinvil, Xi Ai, and Lun-Wei Ku. 2026. Beyond many-shot translation: Scaling in-context demonstrations for low-resource machine translation. *arXiv preprint arXiv:2602.04764*.

Semenov, Kirill, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China, November. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August. Association for Machine Translation in the Americas.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July. Association for Computational Linguistics.

Tezcan, Arda, Bram Bulté, and Bram Vanroy. 2021. Towards a better integration of fuzzy matches in neural machine translation through data augmentation. *Informatics*, 8(1):7.

Tezcan, Arda, Alina Skidanova, and Thomas Moerman. 2024. Improving fuzzy match augmented neural machine translation in specialised domains through synthetic data. *The Prague Bulletin of Mathematical Linguistics*, 122(122):9–42.

Valdez, Susana and Ana Guerberof-Arenas. 2025. "Google Translate is our best friend here": A vignette-based interview study on machine translation use for health communication. *Translation Spaces*, 14(2):253–276.

Valdez, Susana, Floor van Heeswijk, and Noa Warren. 2025. Machine translation at the hospital: Health-

care professionals' perspectives on use, appropriateness, and policy. *Revista Tradumàtica: translation technologies*, (23):244–265.

Vieira, Lucas Nunes. 2024. Machine translation and migration. In Maher, Brigid, Loredana Polezzi, and Rita Wilson, editors, *The Routledge Handbook of Translation and Migration*, pages 221–234. Routledge, London.

Wang, Wenhui, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online, August. Association for Computational Linguistics.

Wassie, Aman Kassahun, Mahdi Molaei, and Yasmin Moslem. 2024. Domain-specific translation with open-source large language models: Resource-oriented analysis. *arXiv preprint arXiv:2412.05862*.

Xu, Jitao, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online, July. Association for Computational Linguistics.

A Prompt template

The prompt template used for LLM-based translation follows a structured format. In the **0-shot** setting, the prompt consists of a task instruction followed by the source sentence:

```
Translate the source text from
{source_language} to {target_language}.
Source: {source_sentence}
Target:
```

In the k**-shot** setting ($k \geq 1$), the prompt is augmented with k source–target examples retrieved from the TM, either through FM retrieval or random selection:

```
Translate the source text from
{source_language} to {target_language}.
Source: {example_source_1}
Target: {example_target_1}
...
Source: {example_source_k}
Target: {example_target_k}
Source: {source_sentence}
Target:
```

B BLEU and COMET scores

B.1 Shot count analysis

B.2 fuzzy-match vs random selection (TranslateGemma-27b)

B.3 NMT vs LLM comparison

Shots	TP-9b	TG-4b	TG-12b	TG-27b
0	21.65	24.37	25.48	29.28
1	23.84	27.35	29.62	33.25
2	24.81	29.57	31.39	34.57
5	25.94	31.91	34.11	36.19
12	27.72	33.01	35.95	38.41
15	28.08	33.85	36.65	38.86
18	28.64	33.65	37.19	39.15
20	28.61	33.90	36.94	38.97

Table 5: Average BLEU score by shot count and model – TowerPlus-9b (TP-9b), TranslateGemma-4b (TG-4b), -12b (TG-12b), and -27b (TG-27b) – with fuzzy-match retrieval, averaged over 14 target languages.

Shots	TP-9b	TG-4b	TG-12b	TG-27b
0	73.79	80.68	86.77	88.25
1	74.55	83.33	88.07	89.34
2	75.20	84.78	88.31	89.49
5	76.05	85.78	88.86	89.80
12	77.35	86.15	88.91	89.54
15	77.68	85.96	89.12	89.21
18	78.12	85.90	89.22	88.53
20	78.00	85.82	88.85	88.20

Table 6: Average COMET (`wmt22-comet-da`) score by shot count and model – TowerPlus-9b (TP-9b), TranslateGemma-4b (TG-4b), -12b (TG-12b), and -27b (TG-27b) – with fuzzy-match retrieval, averaged over 14 target languages.

Target	0-shot	Fuzzy	Random
Arabic	29.52	**38.38**†	34.29
German	43.87	**51.97**†	48.42
Spanish	44.84	**52.58**$^{-}$	51.53
Farsi	20.72	**33.03**†	30.31
Portuguese	41.71	**57.70**‡	55.34
Romanian	34.42	**42.28***	40.58
Russian	38.40	**47.01**‡	42.28
Turkish	27.42	**36.18**‡	32.34
Armenian	29.79	**39.74**‡	34.30
Albanian	36.49	**42.19**$^{-}$	39.85
Georgian	24.80	**31.78**†	29.02
Pashto	20.39	**34.22**‡	29.61
Somali	15.30	**25.62**‡	20.87
Tigrinya	2.29	**11.30**‡	6.44
Average	**29.28**	**38.86**	**35.37**

Table 7: BLEU scores per target language for 0-shot, fuzzy-match (15 shots), and random selection (15 shots) using TranslateGemma-27b. Best scores are in bold; significance markers ($^{-}$/*/†/‡) denote paired bootstrap test results comparing the best strategy against the second-best per language.

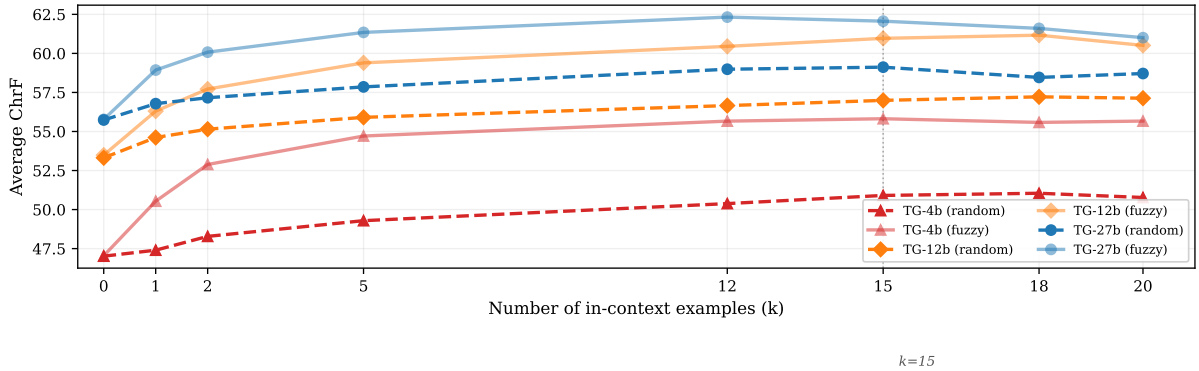

Figure 3: Average ChrF as a function of shot count for the three TranslateGemma (TG) models, TG-4b, TG-12b, and TG-27b, comparing fuzzy-match retrieval (solid lines) with random example selection (dashed lines). TowerPlus-9b is excluded as no random selection data was available.

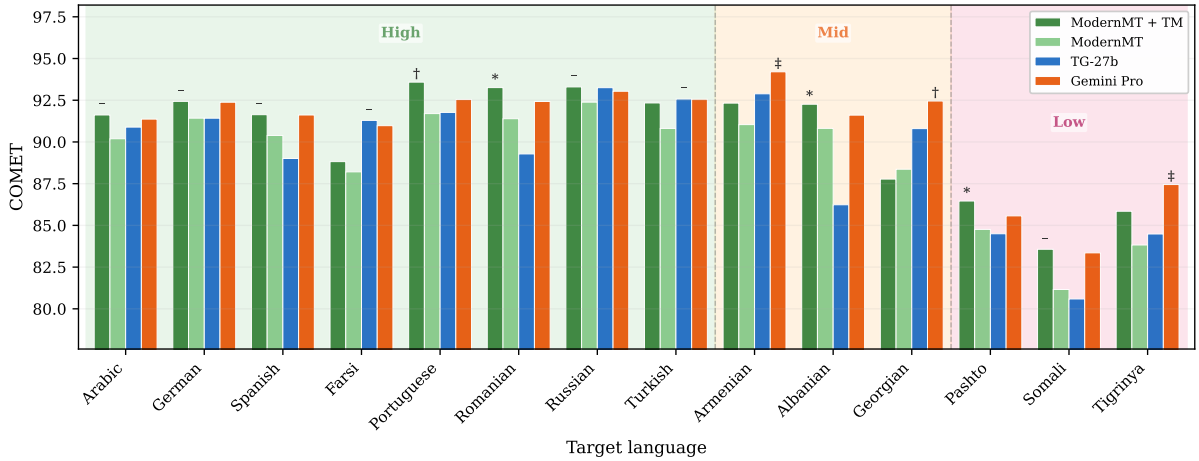

Figure 4: COMET (`wmt22-comet-da`) scores for ModernMT + TM (MMT$^+$), ModernMT (MMT), TranslateGemma-27b (TG-27b), and Gemini Pro (GP) per target language. Languages ordered by resource tier. Significance markers ($^-$/$*$/†/‡) above the best bar denote paired bootstrap test results comparing the best system against the second-best per language.

Target	0-shot	Fuzzy	Random
Arabic	89.35	**90.89**$^-$	90.83
German	91.23	91.42	**91.60**$^-$
Spanish	89.79	89.01	**90.14**$^-$
Farsi	89.18	**91.29**‡	90.25
Portuguese	90.52	**91.77**$^-$	91.62
Romanian	**91.45**$^-$	89.28	91.01
Russian	91.99	**93.25**‡	92.34
Turkish	92.18	92.57	**92.72**$^-$
Armenian	91.91	**92.89**†	91.99
Albanian	**90.32**‡	86.23	88.03
Georgian	89.49	**90.80**$^-$	90.46
Pashto	82.01	**84.50***	83.53
Somali	78.49	**80.59***	79.11
Tigrinya	77.58	**84.48**‡	81.67
Average	**88.25**	**89.21**	**88.95**

Table 8: COMET (`wmt22-comet-da`) scores per target language for 0-shot, fuzzy-match (15 shots), and random selection (15 shots) using TranslateGemma-27b. Best scores are in bold; significance markers ($^-$/$*$/†/‡) denote paired bootstrap test results comparing the best strategy against the second-best per language.

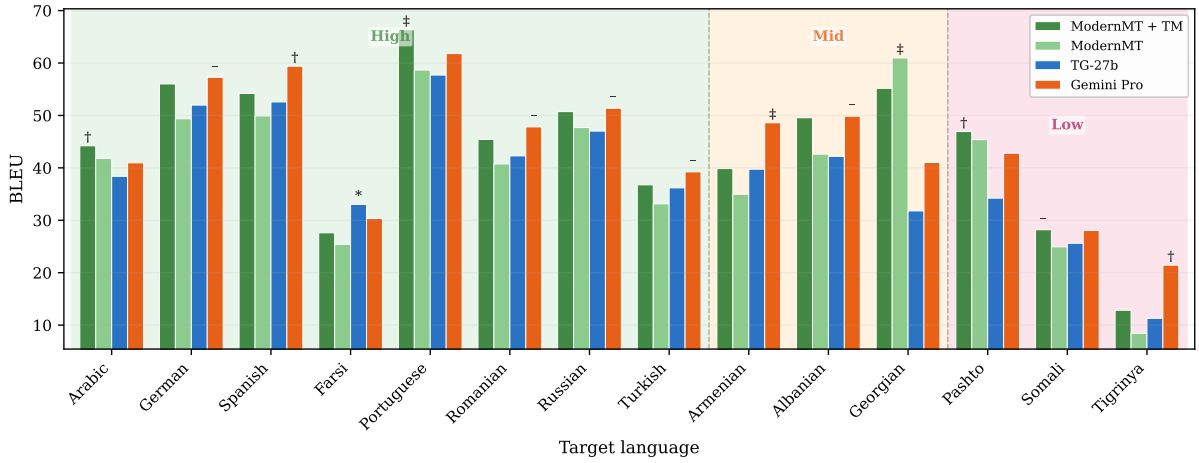

Figure 5: BLEU scores for ModernMT + TM (MMT$^+$), ModernMT (MMT), TranslateGemma-27b (TG-27b), and Gemini Pro (GP) per target language. Languages ordered by resource tier. Significance markers ($^-$/$*$/†/‡) above the best bar denote paired bootstrap test results comparing the best system against the second-best per language.