# Improving Retrieval-Augmented Neural Machine Translation with Monolingual Data

**Maxime Bouthors**[†] **Josep Crego**[†] **Dakun Zhang**[†] **François Yvon**[‡]
[†]SYSTRAN by ChapsVision, 5 rue Feydeau, F-75002 Paris, France
[‡]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
{mbouthors, jcrego, dzhang}@chapsvision.com
yvon@isir.upmc.fr

## Abstract

Conventional retrieval-augmented neural machine translation (RANMT) systems leverage bilingual corpora, e.g., translation memories (TMs). Yet, in many settings, monolingual corpora in the target language are also available. This work explores ways to take advantage of such resources by directly retrieving relevant target language segments, based on a source-side query. For this, we design improved cross-lingual retrieval systems, trained with both sentence level and word-level matching objectives. In our experiments with three RANMT architectures, we assess such cross-lingual objectives in a controlled setting, reaching performances that match those of standard TM-based models. We also showcase our method on two real-world settings, using much larger monolingual corpora, and observe strong improvements over both baseline RANMTs and general-purpose cross-lingual retrievers.

## 1 Introduction

The use of Retrieval-Augmented Generative Models (Li et al., 2022) is rapidly expanding, owing to their built-in ability to condition generations with illustrative segments retrieved from memory. The combination of (machine) retrieval and (human) regeneration has long been a key principle of Computer Aided Translation (CAT) systems (Arthern, 1978; Kay, 1997; Bowker, 2002): segments resembling the source sentence are first retrieved from Translation Memories (TM), providing translators valuable suggestions that can then be edited into a new translation. Such techniques have been transposed in Statistical MT (Koehn and Senellart, 2010), more recently in Retrieval-Augmented Neural Machine Translation (RANMT) e.g., (Gu et al., 2018), making the entire process fully automatic. Most subsequent work has continued to leverage parallel data, performing retrieval in the source language, then using the retrieved target side text in generation. Similar principles underlie few-shot MT with large language models, where both the source and target sides of examples are inserted into the prompt (Brown et al., 2020).

TM-based approaches typically retrieve examples with **lexical fuzzy matchers** (see figure 1 (a)), using BM25 and/or the Levenshtein distance (LED) as filters or rankers (Bouthors et al., 2024). These are known to be computationally efficient and to often surpass continuous-space retrievers (Xu et al., 2020): this is because lexical (source) matches often come with appropriate translation of rare words. However, using source-side lexical similarity raises two issues that may cause inadequate segments to be retrieved: (a) noisy alignments between source and target sides of parallel instances; (b) linguistic divergences between the two languages (Dorr, 1994).

These problems can be sidestepped by performing **retrieval directly on the target side**, as illustrated in figure 1 (b), using Cross-Lingual Information Retrieval (CLIR) techniques (Cai et al., 2021; Tamura et al., 2023). An important additional benefit is to dispense with the use of parallel data and inform the generation process with relevant monolingual resources. As demonstrated by the recurrent use of target side monolingual data thanks to back-translation (Sennrich et al., 2016a),

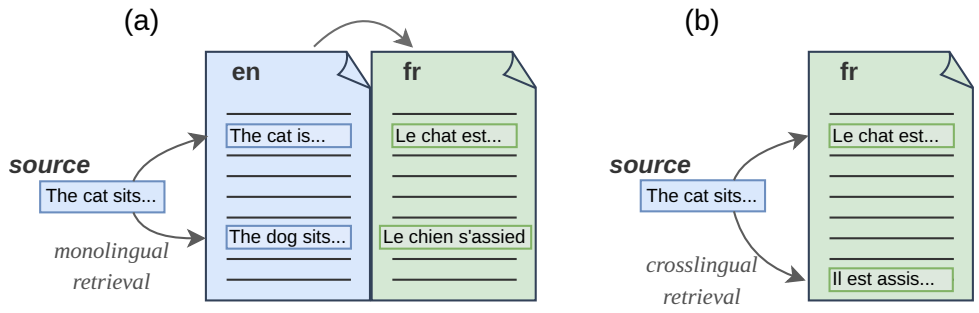


**Figure 1:** Illustration of (a) the classic TM-based retrieval using fuzzy matchers vs. (b) the CLIR-based approach via sentence encoders and cosine similarity.

such resources are much easier to find than parallel ones in many domains, and are also devoid of so-called *translationese phenomena* (Bogoychev and Sennrich, 2020). As such, monolingual corpora constitute a very useful, yet under-studied, source of high-quality segments for RANMT. To search such resources, cross-lingual retrievers are readily available thanks to multilingual sentence encoders such as LASER (Artetxe and Schwenk, 2019b) or LaBSE (Feng et al., 2022). As they only rely on dense representations, such retrievers however often fail to retrieve lexically relevant examples, i.e, examples that would contain translations of the exact source words.

In this work, we propose novel methods to boost the lexical matching abilities of language-agnostic sentence encoders, thereby improving their effectiveness for RANMT. For this, we consider various ways to fine-tune a pre-trained encoder into reproducing the behaviour of lexical fuzzy matchers.[1] We experiment with three language pairs (English-French, German-English and English-Ukrainian) and multiple domains of various sizes and diversities. We consider three types of RANMT systems: an autoregressive source-augmented transformer (Bulte and Tezcan, 2019), a non-autoregressive edit-based model (Bouthors et al., 2023) and a large language model (LLM), with in-context ($k$-shot) learning (Brown et al., 2020). We first study the performance of CLIR techniques *in a controlled setting*, showing that they match standard TM-based models with source-side retrieval. We then *consider much larger setups*, where the target monolingual resources far exceed the amount of parallel data and observe large improvements using our new techniques, which outperform various baseline systems, as well as RANMT architectures relying on generic cross-lingual retrievers.

[1]Code available at https://github.com/Maxwell1447/clir4ramt

## 2 Related Work

**Retrieval Augmented NMT** Our work falls under the scope of Retrieval-Augmented Generation, leveraging both lexical (BM25 (Robertson and Walker, 1994), Levenshtein distance (Levenshtein, 1965)) and semantic methods for the Information Retrieval (IR) component. Semantic methods usually rely on dual encoders trained on parallel data (Gillick et al., 2018) with a contrastive loss (Sohn, 2016). Our focus is on massively multilingual sentence encoders that encode mutual translations as close neighbors, such as LASER (Artetxe and Schwenk, 2019b) and LaBSE (Feng et al., 2022).

In recent years, retrieval-augmented MT systems have received increasing attention. One key motivation is to enhance model transparency by providing users with the retrieved examples that inform the generated output (Rudin, 2019). A simple implementation of this idea encodes the target side of the example(s) along with the source sentence, providing an additional translation context (Gu et al., 2018; Bulte and Tezcan, 2019; Xia et al., 2019; He et al., 2021; Cheng et al., 2022; Agrawal et al., 2023). It is also possible to leverage both the source and target sides of the retrieved example(s) (Pham et al., 2020; Reheman et al., 2023).

Rather than regenerating a complete translation with extended context, edit-based models perform minimal changes to the retrieved examples and generate their output *in a non-autoregressive (NAR) fashion*. Such models are studied in (Niwa et al., 2022; Xu et al., 2023; Zheng et al., 2023), adapting the Levenshtein Transformer (Gu et al., 2019). (2023) generalize this approach to simultaneously handle multiple examples.

MT systems based on Large Language Models (LLMs) can also seamlessly accomodate examples

through in-context learning (ICL), where the LLM prompt is enriched with *demonstrations of the translation task*, comprising both the source and target sides of parallel samples (Radford et al., 2019). Several studies have tried to optimize the performance of such architectures for MT, examining the impact of prompt modifications, the number of demonstrations and the retrieval procedure (Moslem et al., 2023; Vilar et al., 2023; Zhang et al., 2023; Hendy et al., 2023; Bawden and Yvon, 2023; Zebaze et al., 2025).

**Monolingual Data in NMT** Augmenting NMT models with target side monolingual data by back-translating segments is a well-established strategy for improving model training (Sennrich et al., 2016a; Edunov et al., 2018), already considered in RANMT setup (Tezcan et al., 2024). (Cai et al., 2021; Tamura et al., 2023) dispense with back-translation, performing directly the retrieval in the target language with CLIR techniques: the latter uses existing retrievers, while the former jointly trains the retrieval and the MT components.

**Metric Learning** Our lexical alignment objectives (§ 3.2) is a form of neural metric learning: a high-dimensional metric space is reduced to a smaller one (Suárez et al., 2021). This line of work aims to learn distances from the data (Kulis, 2013; Bellet et al., 2015), especially useful for similarity-based IR algorithms (Cakir et al., 2019).

## 3 Method

### 3.1 Problem Formulation

Given a source sentence $\mathbf{x}$ and a corpus of target segments $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, the goal of CLIR is to retrieve $k$ *relevant* examples: $(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_k)$. Our approach closely follows neural-based cross-lingual retrievers, and comprises two main steps. First, $\mathcal{D}$ is encoded with a multilingual sentence encoder $E_\theta$. Then, for any query source sentence $\mathbf{x}$, retrieval is performed as a $k$NN similarity search in the embedding space, using the cosine similarity:

$$\mathrm{sim}_\theta(\mathbf{x}, \tilde{\mathbf{y}}) = \cos(E_\theta(\mathbf{x}), E_\theta(\tilde{\mathbf{y}})). \quad (1)$$

The quality of the retrieved sentences only depends on the choice of $E_\theta$. Usually, $E_\theta$ is trained as a Siamese encoder with a contrastive loss on parallel data (Sohn, 2016), primarily capturing semantic similarity. However, it is desirable for the retrieved $\tilde{\mathbf{y}}_i$ instances to exhibit lexical similarities with the unknown translation $\mathbf{y}$ of the source $\mathbf{x}$. In specialized domains (law, medicine, etc.), adherence to the terminology takes precedence over semantic similarity, thereby motivating the design of our proposed lexical-based alignment objective. To address this need, we study various fine-tuning objectives based on the *Levenshtein similarity*. These objectives are specifically designed **to boost the retrieval of segments with a large numbers of lexical matches in the source query**, by learning a new embedding space, achieved through the optimization of parameters $\theta$, such that:[2]

$$f(\mathrm{sim}_\theta(\mathbf{x}, \tilde{\mathbf{y}})) \approx \mathrm{Lev}(\mathbf{y}, \tilde{\mathbf{y}}), \quad (2)$$

where $f$ is a (potentially learnt) mapping from the codomain of the cosine function into the codomain of Levenshtein similarity function Lev, defined as:

$$\mathrm{Lev}(\mathbf{y}, \tilde{\mathbf{y}}) = 1 - \frac{\Delta(\mathbf{y}, \tilde{\mathbf{y}})}{\max(|\mathbf{y}|, |\tilde{\mathbf{y}}|)}. \quad (3)$$

$\Delta$ is the Levenshtein distance (Levenshtein, 1965).

### 3.2 Training

We consider three objective functions. The first two learn to predict the Levenshtein similarity (3). We choose $f$ to be a learnable function that maps $[-1, 1]$ to $[0, 1]$:

$$f(t) = \sigma(a \times \mathrm{arctanh}(t) + b), \quad (4)$$

with $a$ (slope) and $b$ (position) parameters and $\sigma$ the sigmoid function. The loss is defined as:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \tilde{\mathbf{y}}) = \mathrm{Err}(f(\mathrm{sim}_\theta(\mathbf{x}, \tilde{\mathbf{y}})), \mathrm{Lev}(\mathbf{y}, \tilde{\mathbf{y}})), \quad (5)$$

where the error function Err is either the mean square (MSE) or the mean absolute error (MAE):

$$\mathrm{MSE}(x, y) = (x - y)^2 \quad (6)$$

$$\mathrm{MAE}(x, y) = |x - y|. \quad (7)$$

The third objective is a "learning to rank" objective (Cao et al., 2007). Given a set of $k$ retrieved segments $\tilde{\mathbf{y}}_{[1:k]}$ sorted in decreasing order w.r.t. $\mathrm{Lev}(\mathbf{y}, \cdot)$, it computes:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \tilde{\mathbf{y}}_{[1:k]}) = \sum_{i>j} \max(0, U_{ij}) \quad (8)$$

$$U_{ij} = \mathrm{sim}_\theta(\mathbf{x}, \tilde{\mathbf{y}}_i) - \mathrm{sim}_\theta(\mathbf{x}, \tilde{\mathbf{y}}_j) + m \times |\,\mathrm{Lev}(\mathbf{y}, \tilde{\mathbf{y}}_j) - \mathrm{Lev}(\mathbf{y}, \tilde{\mathbf{y}}_j)|) \quad (9)$$

with $m$ a scaling factor. Equation (9) defines a pairwise margin loss,[3] with an adaptive margin proportional to the absolute difference between the two individual Lev similarities.

[2] As edit distances are only computed monolingually, our method does not assume any lexical overlap between $\mathbf{x}$ and $\tilde{\mathbf{y}}$.

[3] Inspired by the triplet loss (Schultz and Joachims, 2003).

## 4 Data and Metrics

### 4.1 Data

We consider three translation directions: English into French (en–fr), German into English (de-en) and English into Ukrainian (en–uk). For the controlled experiments, we consider a wide variety of textual domains (16), using public parallel datasets (en–fr and de-en) from the Opus corpus (Tiedemann, 2012). Details are in appendix A.

Large scale experiments for English-French are based on a large corpus of Wikipedia pages in French.[4] We split the text in pseudo-sentences via regular expression matching and remove duplicates as well as fuzzy matches from the valid/test sets[5] of the parallel Wikipedia corpus, ending up with about 45M segments. Large-scale experiments with legal texts translated from English into Ukrainian, as in (Tezcan et al., 2024), are in Appendix I .

### 4.2 Metrics

The success of cross-lingual retrieval is assessed with the average target side Levenshtein similarity (eq. (3)) between the 1-best match and the reference, for sentences in the validation set, with retrieval performed from the entire train set. We also report `xsim` error rates, which measure the abililty to identify matched sentences in a bilingual parallel corpus (Artetxe and Schwenk, 2019a).

We assess translation quality with `BLEU` (Papineni et al., 2002) computed with SacreBLEU (Post, 2018),[6] as well as `COMET-20`[7] (Rei et al., 2020).

## 5 Experimental Settings

### 5.1 Retrieval Settings

Our experiments consider the following retrieval methods, with up to $k = 3$ closest examples:

- **`fuzzy-src`**: *source-to-source* fuzzy matching ranked by their Lev similarity, as in baseline RANMT settings;
- **`fuzzy-gold`**: *target-to-target* fuzzy matching, using the same procedure as **`fuzzy-src`**. This retriever requires access to reference target translations as query and *corresponds to an oracle setting*, giving an empirical upper-bound of what an ideal system could achieve. It is only reported for the controlled experiments of § 6.1 and § 6.2;
- **`fuzzy-bt`**: the target side of training segments are back-translated using NLLB (Costa-jussà et al., 2024)[8] into the source language (Sennrich et al., 2016a), then used for retrieval as in **`fuzzy-src`**. This simulates an alternative to cross-lingual retrieval in situations where only target side examples are available;[9]
- **`dense`**: a BERT base (Devlin et al., 2019) model for *source-to-target* retrieval. It is trained from scratch with a contrastive loss (Sohn, 2016). The loss $\mathcal{L}(\mathcal{B})$ for batch $\mathcal{B} = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$ is defined as:

$$-\sum_i \log \frac{\exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{y}_i))}{\sum_j \exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{y}_j))}; \qquad (10)$$

- **`dense+bow`**: we add a bag-of-word loss to **`dense`** as in (Cai et al., 2021) to enforce lexical information in cross-lingual encodings. Given a pair of parallel sentences $(\mathbf{x}, \mathbf{y})$, the loss $\mathcal{L}(\mathbf{x}, \mathbf{y})$ is computed as:

$$-\sum_{w \in \mathcal{Y}} \log p_{\theta_1}(w|\mathbf{x}) - \sum_{w \in \mathcal{X}} \log p_{\theta_2}(w|\mathbf{y}), \qquad (11)$$

  where $\mathcal{X}$, $\mathcal{Y}$ are the bag-of-words for $\mathbf{x}$ et $\mathbf{y}$; $p_{\theta_1}$ and $p_{\theta_2}$ are distinct linear projections of $E_\theta(\mathbf{x})$ and $E_\theta(\mathbf{y})$;
- **`LASER`**: we use LASER (Artetxe and Schwenk, 2019b) as a pre-trained *source-to-target* retriever. Derived from a multilingual RNN translation model, LASER repurposes source embeddings to define multilingual sentence similarity.
- **`LaBSE`**: Similar to **`dense`**, LaBSE (Feng et al., 2022) uses a dual encoder model based on the BERT base architecture and is trained with a contrastive loss; it additionally benefits from pre-training on large sets of monolingual data, and covers more than 110 languages;
- **`ft-[model]-[Err]`**: we fine-tune either one of the aforementioned encoders with the losses of §3.2; Err is one of {MSE, MAE, Rank} (e.g. **`ft-LaBSE-MSE`**).

[4] https://huggingface.co/datasets/Plim/fr_wikipedia_processed.

[5] Each best match with Lev $> 0.9$ is simply removed.

[6] signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`;

[7] Unbabel/wmt20-comet-da, using the defaults settings.

[8] We used `NLLB-200-distilled-1.3B` out-of-the-box. https://huggingface.co/facebook/nllb-200-1.3B

[9] Back-translation is further discussed in § G.

In our setting, fine-tuning updates all the parameters. The global learning rate is 1e-4; as for $a$ and $b$ (eq. 4) it is set to 1e-2. Each source $\mathbf{x}$ is presented with three $\tilde{\mathbf{y}}$ (eq. 5) determined by the top 3 scoring matches obtained with **dense+bow**. The validation score is an in-batch normalized discounted cumulative gain (NDCG) score (Wang et al., 2013).

Retrieval is always performed in an in-domain setting, ensuring that translations for a specific domain are exclusively based on examples from that domain. For all methods, a threshold is further applied to filter out low similarity segments. For each retrieval method, we adjust this threshold on the validation set to achieve an averaged retrieval rate[10] of 50%. A fixed retrieval rate prevents bias toward systems retrieving more examples, making comparisons depend *solely on TM match quality*. Refer to appendix B for a discussion of search configurations.

### 5.2 NMT Architectures

We consider three RANMT architectures. The first is an in-house implementation of NFA (Bulte and Tezcan, 2019), which extends a basic encoder-decoder MT architecture by prepending the target side of TM examples to the source text on the encoder side. As is standard, the retrieved examples used in training are selected based on their Levenshtein distance to the source text; at most one similar example is used. This model thus also has the ability to perform inference from scratch, without using any retrieved example.

The second architecture implements an edit-based, non-autoregressive encoder-decoder model that has the ability to jointly edit up to $k$ examples to generate a translation, reproducing the multi-Levenshtein transformer (TM$^k$-LevT) (Bouthors et al., 2023). Retrieved examples used in training are also selected based on source-side similarity measures; we use up to $k = 3$ examples.

The third architecture is EuroLLM-9B (Martins et al., 2025), specifically designed and trained for multilingual tasks such as Machine Translation. We fine-tune the EuroLLM-9B base model on the en–fr and de-en training corpora using a prompt *that only contains relevant examples in the target language*. A discussion of the adaptation of EuroLLM for this setup is reported in Appendix D.

[10]Fraction of sentences with at least one retrieved example.

| | test en–fr | | test de-en | |
|---|---|---|---|---|
| | Lev ↑ | xsim ↓ | Lev ↑ | xsim ↓ |
| **fuzzy-gold** (oracle) | 50.0 | - | 53.4 | - |
| **fuzzy-src** | 36.0 | - | 21.9 | - |
| **fuzzy-bt** | 31.1 | - | 21.8 | - |
| **dense** | 30.4 | 10.8 | - | - |
| **dense+bow** | 32.7 | 8.0 | 30.9 | 25.2 |
| **ft-dense+bow-MSE** | 34.9 | 7.6 | 33.6 | 25.2 |
| **ft-dense+bow-MAE** | 35.2 | 8.0 | 32.9 | 25.3 |
| **ft-dense+bow-Rank** | 29.9 | 11.9 | 25.9 | 41.5 |
| **LASER** | 32.8 | 10.0 | 33.2 | 26.2 |
| **LaBSE** | 33.8 | 7.4 | 34.4 | **21.0** |
| **ft-LaBSE-MSE** | 36.7 | 7.6 | 34.0 | 24.6 |
| **ft-LaBSE-MAE** | **37.0** | **6.4** | **34.5** | 23.5 |
| **ft-LaBSE-Rank** | 36.0 | 7.3 | 34.2 | 23.3 |

**Table 1:** Average retrieval scores (xsim error and Levenshtein similarity), averaged accross domains.

## 6 Results

We first run controlled experiments serving two main purposes: (a) to compare the quality of various retrieval models (§ 6.1); (b) to evaluate the gap between the baseline source-side retrieval and our proposed CLIR-based approaches (§ 6.2), an analysis that crucially requires parallel datasets. These experiments use two language pairs (en–fr and de-en). This configuration is somewhat ideal, as CLIR methods are primarily meant for cases where only monolingual data are available, as illustrated by the experiments in § 6.3.

### 6.1 Retrieval Scores

Table 1 stores the retrieval scores (§4.2) for the small-scale experiments. For en–fr, source-side fuzzy matching yields more similar segments than CLIR techniques, a gap that vanishes after fine-tuning with MAE for **dense+bow** and even more so for **LaBSE**. For de-en, the gap between the source-side fuzzy matching baseline and the oracle condition is much larger than for en–fr, which might be due to residual misalignments or to morphological differences between German and English.[11] All CLIR techniques widely outperform the monolingual settings (**fuzzy-src** and **fuzzy-bt**). For this language pair, however we do not see much gain from fine-tuning, perhaps owing to a smaller fune-tuning dataset. As expected, fine-tuning with a lexical loss is always slightly detrimental to the

[11]Inflectional processes in German create multiple variants for each lexeme, making exact matches at the word level much less likely in German than in English.

sentence-level retrieval scores (xsim error) for both **dense+bow** and **LaBSE**. Training **dense+bow** along with a bag-of-word loss (eq. (11)) effectively improves the retrieval scores.

### 6.2 Translation Scores

**Retrieval-free Baselines** We use NLLB, NFA and EuroLLM without any retrieved examples as machine translation baselines. Their corresponding BLEU scores, averaged across domains, are respectively 39.0 (en–fr) and 33.8 (de-en) for NLLB and 48.0 (en–fr) and 38.2 (de-en) for NFA. Details per domain are in Table 4.

**TM$^3$-LevT** Translation results for TM$^3$-LevT are in Table 2 (left part). The difference between the oracle (**fuzzy-gold**) and the baseline (**fuzzy-src**) gives an idea of the quality of the monolingual matches: the quality is pretty high in en–fr, owing to the cleaning procedure used for these data, more noisy in de-en, with a gap of about 2.5 BLEU points. In comparison, using artificial back-translated source queries performs much worse than using natural texts. This difference is much larger than what was expected from the retrieval scores of Table 1.

For both language pairs, using cross-lingual search enables to obtain results that compare to source-side fuzzy matching. For en–fr, the best scores in our setting use fine-tuned versions of **LaBSE**, closing the gap in BLEU with **fuzzy-src**, and delivering a small COMET gain. Similar to the retrieval scores (Table 1), we do not observe such benefit of fine-tuning for de-en, where cross-lingual retrieval with **LASER** yields the best overall results, yielding a 1.6 BLEU points improvement over the pure source-side matching.

**NFA** The corresponding results for NFA are in Table 2 (middle part). This model is a much stronger baseline than TM$^3$-LevT and the progress margins (with respect to **fuzzy-gold**) is dimmer. Yet, we see that (a) retrieving examples directly on the target side (with **dense**, **LASER**, and **LaBSE**) improves the example-free baseline by about 3 BLEU points; (b) further fine-tuning the cross-lingual retrievers (with MSE and MAE) fully closes the gap with source-side matching for both metrics and language pairs. We observe that back-translating the monolingual target data also improves baseline systems, albeit by a much smaller margin.

**EuroLLM** Results for the fine-tuned version of EuroLLM are in the right part of Table 2. This system turns out to be the most effective system overall, slightly outperforming NFA with $k = 3$ TM examples. A first observation is that this model is particularly strong for translating into English (from German) where using a CLIR outperforms all baselines. For this model, the averaged gains achieved by fine-tuning the cross-lingual retrievers are very small, and vary accross domains (details in Table 5). For this system, the gap between the **LaBSE** system and the oracle condition (**fuzzy-gold**) is already small in the two language pairs, showcasing the very strong built-in cross-lingual abilities of this model; yet, the difference that remains suggests that lexical similarity should still matter.

**Per-domain analysis** Per-domain BLEU scores are in Tables 3 (TM$^3$-LevT), 4 (NFA), and 5 (EuroLLM). The domains for which we expect CLIR methods to work well need to satisfy two conditions: (a) a large enough set of examples to retrieve from, so that the choice the retriever actually makes a difference; (b) a small diversity of texts in that domain, so that retrieved instances are sufficiently similar to the input. This is well reflected for instance in Table 3 (TM$^3$-LevT) where CLIR techniques (with fine-tuning in en–fr, without in de-en) match the performance of the baseline in all domains, and even outpeform it in specialized domains such as ECB (finance), JRC (law), KDE (information technology) and Wikipedia. We do not see the same gains in Ubuntu (very small TM) or Europarl (large TM, with a diverse set of texts).

For NFA, we also observe that CLIR techniques yield scores that match the baseline; for both mono- and cross-lingual retrieval, the gap to **fuzzy-gold** is small across domains, except Koran (de-en), where CLIR narrows it more effectively.

Regarding EuroLLM, we note that performing CLIR with the default version of **LaBSE** already yields near optimal performance for many domains. Fine-tuning has here a reduced effect on BLEU scores, sometimes slightly improving (e.g., for en–fr: EMEA, KDE, PHP, Wiki, Ubu) or, more rarely, decreasing (e.g., for en–fr, ECB, JRC) performance. These variations highlight the fact that optimal translation scores ultimately also depend on the retrieval pool and the associated example quality.

**Comparison of fine-tuning losses** The standard contrastive loss of eq. (10), when applied in MT,

| | TM$^3$-LevT | | | | NFA | | | | EuroLLM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | test en–fr | | test de-en | | test en–fr | | test de-en | | test en–fr | | test de-en | |
| | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| no example | - | - | - | - | 48.0 | 87.1 | 38.2 | 79.8 | 44.4 | 86.3 | 40.7 | 82.6 |
| **fuzzy-gold** (oracle) | 45.3 | 51.7 | 31.3 | -8.6 | 52.2 | 87.8 | 44.7 | 82.2 | 52.6 | 88.1 | 46.8 | 83.9 |
| **fuzzy-src** | 43.8 | **48.4** | 28.8 | -10.3 | **51.3** | **87.5** | 41.9 | **81.7** | **51.6** | **87.8** | 41.6 | 83.0 |
| **fuzzy-bt** | 40.5 | 43.8 | 22.0 | -22.9 | 49.9 | 87.2 | 38.9 | 81.3 | 46.4 | 86.9 | 41.8 | 82.9 |
| **dense** | 41.5 | 42.2 | - | - | 49.9 | 87.1 | - | - | 50.4 | 87.4 | - | - |
| **dense+bow** | 42.4 | 44.1 | 28.3 | -11.6 | 50.5 | 87.3 | 40.0 | 80.9 | 51.1 | 87.5 | 44.1 | 83.0 |
| **ft-dense+bow-MSE** | 43.1 | 46.0 | 29.0 | -12.7 | 51.0 | **87.5** | 41.6 | 81.5 | 51.3 | **87.8** | 45.1 | **83.3** |
| **ft-dense+bow-MAE** | 43.2 | 46.3 | 29.1 | -12.6 | 51.0 | 87.4 | 41.8 | 81.4 | 51.5 | 87.7 | **45.2** | 83.4 |
| **ft-dense+bow-Rank** | 42.1 | 44.3 | 26.0 | -17.9 | 50.7 | 87.3 | 41.8 | 80.9 | 50.4 | 87.5 | 43.6 | 83.1 |
| **LASER** | 42.5 | 44.3 | **30.4** | **-7.7** | 49.9 | 87.3 | 41.1 | 81.3 | 50.9 | 87.4 | 44.9 | 83.1 |
| **LaBSE** | 43.1 | 45.5 | 29.8 | -9.5 | 50.5 | 87.3 | 41.5 | 81.4 | 51.5 | 87.6 | 45.1 | 83.2 |
| **ft-LaBSE-MSE** | *43.7 | *47.8 | 27.7 | -12.5 | *51.2 | *87.5 | 41.6 | *81.6 | **51.6** | **87.8** | 44.2 | 83.2 |
| **ft-LaBSE-MAE** | *43.6 | *47.7 | 29.1 | -9.2 | *51.2 | *87.5 | *42.0 | 81.5 | 51.5 | **87.8** | 44.9 | **83.3** |
| **ft-LaBSE-Rank** | *43.3 | *47.1 | 28.4 | -12.2 | *51.1 | *87.5 | *41.8 | *81.6 | 51.3 | 87.7 | 44.8 | 83.2 |

**Table 2:** Average translation scores: TM$^3$-LevT, NFA, EuroLLM, all domains. Significant results (p=0.05) w.r.t. **LaBSE** are marked with *. Best CLIR results are in **bold**. When **fuzzy-src** is not outperformed, its score is in **bold**.

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| **fuzzy-gold** | 58.5 | 64.8 | 32.5 | 61.3 | 60.5 | 44.3 | 27.0 | 41.0 | 31.2 | 40.5 | 36.9 | 32.1 | 8.2 | 48.7 | 47.6 | 19.8 |
| **fuzzy-src** | 56.7 | **62.2** | **31.9** | **58.9** | 58.8 | 41.5 | **27.2** | **40.3** | **30.7** | **38.9** | 34.6 | 29.2 | 8.7 | 45.8 | 43.3 | 16.9 |
| **fuzzy-bt** | 51.0 | 54.7 | **31.9** | 51.3 | 55.1 | 35.2 | 27.0 | 38.7 | **30.7** | 36.2 | 33.3 | 21.3 | 4.5 | 27.5 | 39.2 | **17.5** |
| **LASER** | 56.5 | 60.5 | 31.7 | 56.0 | 58.6 | 36.6 | 26.9 | 39.7 | **30.6** | 36.4 | 33.9 | **29.3** | 11.3 | 48.9 | **46.6** | 16.2 |
| **LaBSE** | **57.0** | 59.8 | 31.3 | 57.8 | **59.4** | 40.8 | 26.9 | 39.8 | 30.5 | 37.0 | 33.9 | 29.0 | 9.3 | **49.2** | 45.4 | 16.2 |
| **ft-LaBSE-MSE** | **57.0** | **62.0** | **31.8** | 58.6 | 58.6 | **42.5** | **27.0** | 39.9 | 30.5 | **38.1** | **34.9** | 28.1 | 10.9 | 41.4 | 41.1 | 17.0 |
| **ft-LaBSE-MAE** | 56.9 | 61.4 | 31.6 | **58.7** | 58.7 | **42.5** | **27.0** | 39.9 | 30.4 | 37.7 | 34.5 | 28.7 | **11.7** | 44.4 | 43.7 | 17.0 |
| **ft-LaBSE-Rank** | 56.6 | 61.1 | 31.7 | 57.7 | 58.0 | 41.7 | 26.9 | **40.1** | 30.6 | 37.6 | 34.1 | 28.7 | 9.6 | 44.6 | 42.5 | 16.6 |

**Table 3:** Per domain BLEU scores for TM$^3$-LevT systems. Contaminated domains for NLLB (used in back-translation) are underlined.

primarily seeks to identify parallel sentences within collections of monolingual segments. Yet, RANMT requires examples that are lexically similar to the source sentences. The bag-of-word loss of eq. (11) effectively increases the rate of lexically similar sentences, thus enhancing the translation scores. As for our three fine-tuning strategies, **MSE** and **MAE** (eq. (5)) yield comparable performances. The per-domain analysis further highlights their consistent behavior. As for **Rank** (eq. (9)), we observe that the rank-based loss consistently underperforms the other two, with a few exceptions (PHP for NFA; PHP, Wiki and KDE (de-en) for TM$^3$-LevT).

**On data contamination** We identify two sources of data contamination. First, NFA training data partly overlaps with some of our test sets, in varying proportion (from 0 to 100) across domains (see Table 4). The intuition is that such contamination should make this model less sensitive to the choice of good examples. In the per-domain results of Table 4, this sensitivity is assessed by the gap between **example-free** and **fuzzy-gold** scores. As this gap is large for most domains (except Europarl and News in en–fr, and Sub in de-en), we can still clearly assess the benefits of cross-lingual retrieval for the large majority of domains. Second, the training data of NLLB, used for back-translation, also partly overlaps with our test, which may boost the efficiency of the **fuzzy-bt** baseline. Even for these domains (e.g., ECB, EMEA, JRC-Acquis), cross-lingual methods still achieve much higher BLEU scores than back-translation.

### 6.3 Large-scale Experiments

**Retrieving in a monolingual corpus** We use the models trained on the en–fr[12] dataset to retrieve segments from the large monolingual Wikipedia dataset (see §4.1), assessing performance on the Wikipedia test set previously used in the controlled

[12]A similar experiment, targeting translation from English into Ukrainian is in Appendix I; these results also highlight the benefits of CLIR over monolingual retrieval in parallel data, especially when using a fine-tuned version of **LaBSE**.

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| contam. rate | 33.9 | 58.3 | 8.9 | 0 | 2.1 | 6.8 | 100 | 6.8 | 29.8 | 0 | 0 | 47.5 | 0 | 54.0 | 17.3 | 8.9 |
| NLLB | 44.5 | 40.3 | 35.9 | 39.6 | 50.7 | 34.2 | 32.1 | 39.0 | 41.0 | 35.9 | 35.4 | 29.1 | 23.5 | 46.4 | 38.8 | 31.4 |
| no example | 58.9 | 55.9 | 39.9 | 48.8 | 62.3 | 42.8 | 50.1 | 45.6 | 43.9 | 43.5 | 36.4 | 39.5 | 13.2 | 55.5 | 51.6 | 31.0 |
| **fuzzy-gold** | 65.2 | 65.1 | 40.3 | 59.7 | 68.1 | 46.6 | 50.1 | 48.1 | 44.1 | 47.8 | 39.0 | 45.6 | 21.9 | 63.9 | 60.5 | 31.7 |
| **fuzzy-src** | **64.5** | **63.0** | **40.0** | 58.4 | **66.6** | 45.2 | 50.0 | 47.3 | 44.0 | **46.8** | **38.3** | **44.3** | 14.2 | **62.3** | 57.4 | 31.1 |
| **fuzzy-bt** | 62.4 | 60.4 | 39.8 | 54.7 | 64.9 | 44.5 | **50.1** | 46.8 | 43.7 | 44.7 | 37.0 | 40.1 | 13.2 | 55.6 | 54.8 | 31.1 |
| **LASER** | 63.7 | 61.6 | 39.8 | 56.8 | 65.4 | 44.0 | 50.0 | 46.8 | **44.1** | 46.1 | 37.1 | 41.4 | 16.1 | 60.7 | 56.0 | 31.1 |
| **LaBSE** | 63.8 | 62.5 | 39.0 | 57.7 | 65.0 | 45.3 | 50.0 | 46.9 | **44.1** | 46.0 | 37.7 | 43.0 | 15.6 | 60.6 | 56.9 | 31.1 |
| **ft-LaBSE-MSE** | **64.3** | **63.0** | **40.0** | 58.2 | **66.6** | 45.3 | **50.1** | **47.5** | 44.0 | **46.8** | 37.9 | 43.1 | 16.8 | 60.3 | 56.9 | **31.2** |
| **ft-LaBSE-MAE** | 64.2 | **63.0** | 39.9 | **58.5** | **66.6** | **45.5** | **50.1** | 47.4 | 43.9 | 46.3 | 37.8 | 43.3 | **17.1** | **61.4** | **57.5** | 30.9 |
| **ft-LaBSE-Rank** | 64.1 | 62.6 | 39.8 | 58.1 | 66.3 | 45.3 | 50.0 | **47.5** | 44.0 | 46.0 | **38.1** | **43.7** | 15.9 | 61.0 | 57.2 | **31.2** |

**Table 4:** Per domain BLEU scores for NLLB and NFA systems. Contaminated domains for NLLB (used in back-translation) are underlined. The contamination rate (top line) for NFA is the percentage of overlap of its training data with the test set.

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| no example | 49.9 | 47.7 | 39.2 | 45.7 | 55.5 | 40.9 | 34.0 | 48.1 | 42.0 | 41.3 | 44.2 | 41.6 | 23.3 | 54.4 | 52.9 | 31.1 |
| **fuzzy-gold** | 63.7 | 68.4 | 39.5 | 65.2 | 66.2 | 51.1 | 33.8 | 52.1 | 41.9 | 48.9 | 48.2 | 50.4 | 25.6 | 63.6 | 62.2 | 32.2 |
| **fuzzy-src** | 62.6 | **66.9** | **39.5** | **63.3** | 64.9 | 49.0 | 33.7 | 51.6 | **41.7** | **47.4** | 47.0 | 43.3 | 23.6 | 55.1 | 54.3 | 31.8 |
| **fuzzy-bt** | 52.0 | 54.1 | 39.3 | 49.1 | 56.5 | 43.2 | **33.8** | 50.4 | 41.5 | 45.0 | 46.1 | 42.5 | 23.6 | 54.8 | 56.4 | **31.9** |
| **LASER** | 63.2 | 65.3 | 39.4 | 62.1 | 65.5 | 46.1 | **33.8** | 51.0 | **41.7** | 45.8 | 46.4 | 46.6 | **24.4** | 62.4 | 59.9 | 31.4 |
| **LaBSE** | **63.5** | 66.1 | 39.4 | **63.3** | **65.8** | 48.4 | **33.8** | 51.2 | 41.6 | 46.4 | 46.5 | 47.2 | 23.7 | **62.7** | **60.3** | 31.4 |
| **ft-LaBSE-MSE** | 62.8 | 66.3 | 39.4 | 63.2 | 64.9 | **50.0** | **33.8** | **51.7** | **41.7** | 46.8 | **47.3** | 46.6 | 24.1 | 60.0 | 58.7 | 31.5 |
| **ft-LaBSE-MAE** | 62.8 | 65.9 | 39.4 | **63.3** | 65.1 | 49.6 | **33.8** | 51.5 | **41.7** | 46.8 | 47.1 | **47.6** | 24.2 | 61.3 | 59.9 | 31.6 |
| **ft-LaBSE-Rank** | 62.4 | 65.3 | 39.4 | 62.7 | 64.6 | 49.3 | **33.8** | 51.5 | **41.7** | **47.3** | 46.7 | 47.5 | 23.8 | 61.1 | 60.2 | 31.5 |

**Table 5:** Per domain BLEU scores for EuroLLM.

experiments. We use the same filtering threshold as for the controlled experiments to keep the settings comparable and ensure that the only difference is the increased size of the retrieval set. We only consider variants of **LaBSE**, which achieve the best results in the small data condition.

Wikipedia is a challenging domain, due to the variety of topics that are addressed in the encyclopedia and the high level of formality of the text. NLLB already represents a strong baseline for that domain. Our results are in Table 6. The first observation is that increasing the retrieval set (from 200K segments in the controlled experiments – Tables 3 and 4 – to 45M segments) yields significant performance boosts across the board (e.g, +1.6 BLEU for NFA and **fuzzy-bt**, +0.4 BLEU points for NFA and **LaBSE**). With further lexical fine-tuning, the benefits of CLIR are again very clear, with large BLEU differences relative to the small data condition:+3.9 for **ft-LaBSE-MAE** with TM$^3$-LevT, +3.4 when using NFA, +2 using EuroLLM.

The effect of the size of the retrieval pool are discussed in Appendix H.

We also monitor in Table 7 two additional scores: the retrieval rate (percentage of segments for which at least one example is found) and the average Lev similarity between the reference and its closest retrieved neighbor. For all retrieval methods, we see

| | TM$^3$-LevT | | NFA | | EuroLLM | |
|---|---|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| no example | - | - | 36.4 | 84.8 | 44.2 | 86.3 |
| **fuzzy-bt** | 35.6 | 76.7 | 38.6 | 84.8 | 47.1 | 86.4 |
| **LaBSE** | 37.7 | 77.2 | 40.1 | **85.0** | 48.8 | **86.5** |
| **ft-LaBSE-MSE** | 38.3 | 77.2 | 40.9 | 84.8 | 49.0 | 86.3 |
| **ft-LaBSE-MAE** | **38.7** | **77.6** | **41.2** | 84.9 | **49.2** | 86.3 |
| **ft-LaBSE-Rank** | 38.5 | 77.5 | 40.8 | **85.0** | 48.6 | 86.3 |

**Table 6:** BLEU and COMET scores for TM$^3$-LevT, NFA and EuroLLM with the large French Wikipedia retrieval pool.

large increases, meaning a larger number of examples is retrieved and that their average quality (i.e, their similarity to the reference) also improves, highliting the usefulness of monolingual data for this domain.

The translation examples of Appendix F illustrate the benefits of enlarging the retrieval pools, notably because they provide longer lexical matches, thereby improving the generated text.

## 7 Conclusions and Outlook

In this paper, we have explored various approaches to take advantage of monolingual examples in retrieval-augmented neural machine translation. Our main conclusion, resulting from experiments on 3 language directions and 16 varied textual domains, is that retrieving directly in the target lan-

| | TM | | mono | |
|---|---|---|---|---|
| | RR ↑ | Lev ↑ | RR ↑ | Lev ↑ |
| `fuzzy-bt` | 39.9 | 24.0 | 60.2 | 31.0 |
| `LaBSE` | 30.9 | 25.6 | 59.6 | 35.2 |
| `ft-LaBSE-MSE` | 33.8 | 27.2 | 77.7 | 37.7 |
| `ft-LaBSE-MAE` | 31.3 | 27.5 | 59.0 | 38.1 |
| `ft-LaBSE-Rank` | 25.7 | 26.7 | 61.2 | 36.7 |

**Table 7:** Comparison of the retrieval rates (RR) and average Levenshtein similarity (Lev) on the Wikipedia test set of the systems retrieving in the TM vs. in the large monolingual corpus (mono).

guage can be as effective as fuzzy-matching on the source-side, especially when retrievers have been fine-tuned with lexical matching objectives. Using these techniques, we were also able to obtain large BLEU increases over strong baselines on the Wikipedia dataset (up to +3.8 BLEU points for our best configuration), showcasing the benefits of RANMT in a large-scale setup. In our implementation, cross-lingual retrieval is no more costly than performing fuzzy matches in the source language.

We identify several ways to improve these results: (a) by fine-tuning the retrieval models on very large multilingual datasets, adopting data settings that are comparable to the pre-training of LASER/LaBSE; (b) using more recent sentence embedders such as SONAR (Duquenne et al., 2023); (c) using CLIR techniques to train the RANMT models, instead of relying on source-side matches as we have done here; (d) revisiting the other aspect of the retrieval pipeline – e.g., relaxing the filtering threshold during example selection.

## Limitations

In this work, we have focused on extending a translation setting that is common in the translation industry, where translation memories have repeatedly been shown to speed up the translation process and improve the consistency of the resulting texts. In such setting, it is also common to have access to large monolingual resources, as has been studied eg., by (Tezcan et al., 2024) for the legal domain and for (Tamura et al., 2023) in the scientific domain. By design, this setting is only applicable for so-called "high-resource" language pairs, in which high-quality systems can be built and improved upon using auxiliary resources. In our selection of language pairs, we have favored domain diversity over language diversity, to better highlight how the availability of relevant examples impacts the overall translation quality. As our experiments show, the effectiveness of the cross-lingual approach does not depend on the similarity of the source and target languages, but mostly relies on the quality of cross-lingual sentence alignment. As shown e.g., by (Feng et al., 2022), very precise alignments can be achieved, even for distant language pairs such as English-Chinese.

We have thus deliberately chosen to exclude translation from and into "low-resource" languages, for which many resources and components are still difficult to collect (eg., parallel corpora to train sufficiently good NMT baselines, to be used for back translation; and also training and test data in specialized domains, where the TM-based approach is most likely to retrieve very good matches). We reckon that also improving generic NMT systems that could handle translations for those language pairs is another important research area (see, e.g., (Zebaze et al., 2025)), albeit distinct from what we have chosen to study in this work.

## Ethical Statement

There are no ethical issues with this work.

## Acknowledgments

This work was funded by the French Agence Nationale de la Recherche (ANR) under the project TraLaLaM ("ANR-23-IAS1-0006"). This work was also granted access to the HPC resources of IDRIS under allocation 2025- AD011015117R2 made by GENCI. The authors wish to thank the reviewers for their insightful comments and suggestions.

## References

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

Aharoni, Roee and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online, July. Association for Computational Linguistics.

Artetxe, Mikel and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

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

Arthern, Peter J. 1978. Machine translation and computerised terminology systems - a translator's viewpoint. In Snell, Barbara M., editor, *Translating and the Computer*, London, UK, November 14. Aslib Proceedings.

Bawden, Rachel and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland, June. European Association for Machine Translation.

Bellet, Aurélien, Amaury Habrard, and Marc Sebban. 2015. *Metric learning*. Morgan & Claypool Publishers.

Bogoychev, Nikolay and Rico Sennrich. 2020. Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Bojar, Ondřej and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F. Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July. Association for Computational Linguistics.

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

Bouthors, Maxime, Josep Crego, and François Yvon. 2024. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico, June. Association for Computational Linguistics.

Bowker, Lynne. 2002. *Computer-aided translation technology: A practical introduction*. University of Ottawa Press.

Brown, Tom, 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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Larochelle, H., M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Burlot, Franck and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium, October. Association for Computational Linguistics.

Cai, Deng, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 7307–7318, Online, August. Association for Computational Linguistics.

Cakir, Fatih, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. 2019. Deep metric learning to rank. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Cao, Zhe, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Caswell, Isaac, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In Bojar, Ondřej,

Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, August. Association for Computational Linguistics.

Cheng, Xin, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. Neural machine translation with contrastive translation memories. In Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Costa-jussà, Marta R., 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, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, June. ISBN: 1476-4687 tex.date-added: 2024-08-21 15:32:57 +0200 tex.date-modified: 2024-08-21 15:32:57 +0200.

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark, September. Association for Computational Linguistics.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dorr, Bonnie J. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *CoRR*, abs/2401.08281.

Duquenne, Paul-Ambroise, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *CoRR*, abs/2308.11466.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November. Association for Computational Linguistics.

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

Gillick, Daniel, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *CoRR*, abs/1811.08008.

Gu, Jiatao, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2018. Search Engine Guided Neural Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April.

Gu, Jiatao, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

He, Qiuxiang, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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.

Hendy, Amr, 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. *CoRR*, abs/2302.09210.

Kay, Martin. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1/2):3–23.

Koehn, Philipp and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In Zhechev, Ventsislav, editor, *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA, November 4. Association for Machine Translation in the Americas.

Kulis, Brian. 2013. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364.

Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November. Association for Computational Linguistics.

Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Li, Huayang, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *CoRR*, abs/2202.01110.

Martins, Pedro Henrique, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June. European Association for Machine Translation.

Niwa, Ayana, Sho Takase, and Naoaki Okazaki. 2022. Nearest neighbor non-autoregressive text generation. *CoRR*, abs/2208.12496.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Ammar, Waleed, Annie Louis, and Nasrin Mostafazadeh, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Pham, Minh Quang, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. 2020. Priming neural machine translation. In Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527, Online, November. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Reheman, Abudurexiti, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13519–13527, Jun.

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, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins.

2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Robertson, Stephen and Steve Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th ACM Conference on Research and Development in Information Retrieval (SIGIR), Dublin, Ireland*, pages 232–241, 01.

Rudin, Cynthia. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May.

Schultz, Matthew and Thorsten Joachims. 2003. Learning a distance metric from relative comparisons. In Thrun, S., L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sohn, Kihyuk. 2016. Improved deep metric learning with multi-class n-pair loss objective. In Lee, D., M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Suárez, Juan Luis, Salvador García, and Francisco Herrera. 2021. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322.

Tamura, Takuya, Xiaotian Wang, Takehito Utsuro, and Masaaki Nagata. 2023. Target language monolingual translation memory based NMT by cross-lingual retrieval of similar translations and reranking. In Utiyama, Masao and Rui Wang, editors, *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 313–323, Macau SAR, China, September. Asia-Pacific Association for Machine Translation.

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*, 122:9–42, 12.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Vilar, David, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. 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 15406–15427, Toronto, Canada, July. Association for Computational Linguistics.

Wang, Yining, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In Shalev-Shwartz, Shai and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA, 12–14 Jun. PMLR.

Xia, Mengzhou, Guoping Huang, Lemao Liu, and Shuming Shi. 2019. Graph Based Translation-Memory for Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAAI, pages 7297–7304, Jul.

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

Xu, Jitao, Josep Crego, and François Yvon. 2023. Integrating translation memories into non-autoregressive

machine translation. In Vlachos, Andreas and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1326–1338, Dubrovnik, Croatia, May. Association for Computational Linguistics.

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

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zheng, Kangjie, Longyue Wang, Zhihao Wang, Binqi Chen, Ming Zhang, and Zhaopeng Tu. 2023. Towards a unified training for Levenshtein transformer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June.

# A Parallel Data Specification

The en–fr data is a clean version[13] of the one used by (Xu et al., 2020). Validation and test sets each contain $1,000$ segments. The de-en corpus is borrowed unchanged from (Aharoni and Goldberg, 2020) and is also used in (Cai et al., 2021; Agrawal et al., 2023; Bouthors et al., 2024). The associated validation and test sets contain $2,000$ segments. The per-domain corpus sizes are available in table 8. The en–uk corpus is the one provided by (Tezcan et al., 2024), containing about 286K training sentences, 2000 validation sentences and 1898 test sentences. It corresponds to various legal texts.

Note that the preprocessing stages were performed with *subword-nmt* (Sennrich et al., 2016b) for tokenization and BPE, and *fairseq* (Ott et al., 2019) for binarization.

# B Search Parameterization

For both methods, we search up to $k$=3 examples.

## B.1 Fuzzy Matching

The execution of the fuzzy matching methods (**`fuzzy-src`**, **`fuzzy-bt`** and **`fuzzy-gold`**) is performed using an open source library[14]. During the evaluation, there is no filter applied before scoring the segments with Lev. However, the examples associated to the training samples are selected with Levand a BM25 filter. This ensures a reasonable time to build the set of examples for training.

## B.2 FAISS

The execution of the CLIR techniques is done with library FAISS (Douze et al., 2024). The $k$NN search is performed on GPU (IndexFlatIP), without IVF (for faster search) or quantizer (for lower memory footprint) to ensure the retrieval of the best matches. Indeed, both optimizations can prevent the model from finding the optimal $k$ nearest neighbors.

# C NMT Architectures Configuration

Both `TM`$^3$`-LevT` and `NFA` models rely on the default backbone Transformer architecture (Vaswani et al., 2017). We train one instance of `TM`$^3$`-LevT` for each language pair (en–fr and de-en), using only the training data available in the corresponding datasets. Our `NFA` models are trained on much larger sets of parallel data, including a good share of Web data. We use these models to showcase the effectiveness of cross-lingual retrieval for RAMNT systems trained with large, high-quality, parallel datasets (8M sentences for en–fr and 11.5M for de-en). After inspecting the training datasets of `NFA` models, we found that some of them partly overlap with the training data; the corresponding contamination rates are reported in Table 4.

As we compare retrieval techniques *for fixed MT architectures*, these overlaps do not affect on our main conclusions (see also the discussion in §6.2).

In addition, we use the multilingual NLLB model (Costa-jussà et al., 2024) out-of-the-box for both language pairs.[15] This large encoder-decoder model is used to generate the back-translation data; it also provides a baseline against which to appreciate the performance of RANMT architectures. As for the `NFA` models, the training data of NLLB partly overlaps with our test sets, for domains ECB, EMEA, JRC-Acquis and OpenSubtitles. Contamination is discussed in §6.2.

# D Adapting `EuroLLM`

LLMs are well suited for in-context learning in $k$-shot settings (Radford et al., 2019): a set of

[13] We filter noisy parallel pairs with COMETKiwi (Rei et al., 2022) and prepare a new train/valid/test split.

[14] https://github.com/SYSTRAN/fuzzy-match

[15] version: `NLLB-200-distilled-1.3B` https://huggingface.co/facebook/nllb-200-1.3B

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| size | 149k | 272k | 1,9M | 44k | 492k | 126k | 133k | 7k | 148k | 6k | 597k | 223k | 18k | 467k | 248k | 500k |
| % | 3.8 | 7.0 | 49.0 | 1.1 | 12.7 | 3.3 | 3.4 | 0.2 | 3.8 | 0.2 | 15.4 | 15,3 | 1,2 | 32,1 | 17,0 | 34,3 |

**Table 8:** Size of each domain, and its proportion w.r.t. the size of the corpus of the same language pair.

$k$ demonstrations of the task are presented to the model so that the completion of the task fulfills the prompt pattern/template. In our case, the situation is slightly different as we *only include target-side examples*, which are not suitable demonstrations of the translation task.

To adapt an LLM to this setting, we first compare (a) prompting and (b) fine-tuning strategies on the Wikipedia domain.[16]

Regarding (a), we try to take advantage of the in-context abilities of the *Instruct* version of `EuroLLM-9B`, simulating in-context examples with the follow prompt patterns:

- **naive**:
  *Consider the following **French** similar translations.*
  *{k examples in **French**}*
  *Translate the following text from **English** to **French**.*
  ***English**:*
  *{source text}*
  ***French**:*

- **empty:**
  *Consider the following **English-French** examples.*
  ***English**:*

  ***French**:*
  *{**French** example #1}*
  *[...]*
  *Translate the following text from **English** to **French**.*
  ***English**:*
  *{source}*
  ***French**:*

- **copy:**
  *Consider the following **English-French** examples.*
  ***English**:*
  *{**French** example #1}*
  ***French**:*
  *{**French** example #1}*
  *[...]*
  *Translate the following text from **English** to **French**.*
  ***English**:*
  *{source}*
  ***French**:*

The **naive** prompt gives a simple description of the task, where only the target side of examples are introduced. The **copy** and **empty** prompts simulate a standard $k$-shot prompt, where the missing source text is either left empty or contains a copy of the target text. We compare these prompts with a **0-shot** setting, where no example is provided to the model (starting directly with "*Translate the following text*" for any of the three templates).

Regarding (b), we fine-tune the *base* `EuroLLM-9B` model with the **naive** prompt with up to 20K sentences from the training set for each domain. Fine-tuning is performed with QLoRA (Dettmers et al., 2023) with a $r$=16, $\alpha$=32, dropout=0.05, no bias, applied on all linear layers, and 4-bit quantization. Training is performed with huggingface implementation[17] with a batch size of 32 for a single epoch and a learning rate of 2e-10. The in-context examples used for training are retrieved using the `fuzzy-src` strategy.

| Instruct | | | | **ft**-base |
|---|---|---|---|---|
| 0-shot (k=0) | naive (k=3) | copy (k=3) | empty (k=3) | naive (k=3) |
| 36.5 | 36.5 | 38.4 | 37.5 | **49.5** |

**Table 9:** `BLEU` score of various `EuroLLM-9B` based systems on the Wikipedia test set (0-shot and 3-shot). Target examples are retrieved with `LaBSE` with a custom threshold of 0.6.

We observe the superiority of the fine-tuning strategy compared to the various prompt-based approaches in Table 9. As some of these differences

[16] Another alternative, that we do not consider in this work, would be to use backtranslation to automatically craft the missing source for in-context examples.

[17] https://huggingface.co/docs/transformers/trainer

might be explained by the implicit language and domain adaptation that takes place during model fine-tuning, we additionally fine-tune the base model with a *standard* $k$-shot prompt, including the source and target side of each examples. This setting directly compares with the *instruct* model, the only difference being the adaptation that is performed during fine-tuning; it also allows us to assess the performance loss that happens when using only target side examples. Results are in Table 10 for a mixture of 5 domains.

| EuroLLM model | source | k | BLEU |
|---|---|---|---|
| Instruct | - | 0 | 42.9 |
| | ✓ | 3 | 54.0 |
| **ft**-base (src+tgt) | - | 0 | 46.7 |
| | ✓ | 3 | 58.5 |
| **ft**-base (tgt) | - | 0 | 47.3 |
| | ✘ | 3 | 57.5 |

**Table 10:** Average BLEU score of various EuroLLM-9B system fine-tuning schemes for ECB, EMEA, Europarl, JRC-Acquis and Wikipedia test sets (k-shot). Parallel examples are retrieved with **fuzzy-src** using a custom threshold of 0.

In Table 10, we observe that in fact the domain/language adaptation of EuroLLM accounts for about +4 BLEU points (**ft-**base (src+tgt) vs. instruct). For the sole Wikipedia domain, this adaptation yields about +6 BLEU points improvement. This confirms that method (b) is in fact clearly superior to simple prompting strategies. Therefore, this is the method used for all the results reported in Section 6. Moreover, we observe that the absence of the source side of $k$-shot examples is slightly detrimental to the model performance ($\approx$ -1 BLEU point between the two fine-tuned versions, both for the 0-shot and 3-shot settings), suggesting that the target side of the retrieved examples contains most of the information needed to guide the model towards the desired translation.

## E Computational Cost of Retrieval

The computational cost can be dived into two categories: a fixed cost (independant from the number of sentences to translate) and a variable cost (proportional to this number). The retrieval process for training falls into the former, as well as the encoding of the whole set of monolingual target segments (for CLIR techniques) and the indexation of the TM (for fuzzy matching techniques). The variable costs lie in the fuzzy match search with filtering and the computation of Lev, and in the CLIR search with the encoding of the source sentences, then the $k$NN search. In our experiments, we removed the filter (BM25) so that we obtain the best possible fuzzy matches (**fuzzy-src**, **fuzzy-bt** and **fuzzy-gold**), at the cost of a $\sim$100 times higher latency. However, when applying a BM25 filter to preselect 100 segments, we observe that both fuzzy matching and CLIR techniques can retrieve their closest match in about $\sim$1ms. We obtained this result un corpus ECB (150K sentences) with optimmal conditions:

- the fuzzy matching is performed on a 8 cores CPU, thus parallelizing the search on 8 simultaneous source sentences;
- as for CLIR, source sentences are encoded in batches of size 50, and the $k$NN search is handled by the FAISS library. We used a V100-32GB GPU.

## F Retrieval and Translation Examples

An illustration of the retrieved examples is provided is table 11. It highlights the enhanced capability of **ft-LaBSE-MAE** to retrieve lexically closer examples, rather than semantically close ones. This is particularly clear in the second example, which mentions Shasta, a volcano. While **LaBSE** retrieves topic-related segments, **ft-LaBSE-MAE** tends to select off-topic sentences that nonetheless contain a higher number of shared lexical items.

## G Back-Translation

Back-translation (BT) has long been identified as a very effective technique to handle monolingual data in Statistical Machine-Translation (Bojar and Tamchyna, 2011), then in Neural Machine Translation (Sennrich et al., 2016a; Currey et al., 2017; Edunov et al., 2018; Burlot and Yvon, 2018; Caswell et al., 2019). Assuming the task is to translate from L1 into L2, and that monolingual texts in L2 are available, the BT approach translates these "backward" into L1, to obtain an artificial parallel corpus, pairing automatically generated source sentences with actual human-written target sentences. In NMT, it is custom to jointly use "natural" and "artificial" subsets of data to train and/or adapt translation models. This is illustrated in the first two columns (a) and (b) in Table 12, using no-retrieval ($k = 0$), and corresponding respectively to the default (only

| | retriever | Lev | |
|---|---|---|---|
| reference | | | Le mannequin apparaît dans plusieurs séries télévisées au début des années 2000. (*The model also appears in several TV shows in the early 2000s.*) |
| train | **LaBSE** | 0.14 | Cette série fut diffusée sur TF1 pendant l'été **2000.** (*This show was broadcast on TF1 during the summer of 2000.*) |
| train | **ft-LaBSE-MAE** | 0.23 | Elle est aussi apparue dans des **séries télévisées.** (*She also played in some TV shows.*) |
| mono | **LaBSE** | 0.46 | Elle se fait connaître par sa série **au début des années 2000.** (*She made a name for herself thanks to her show in the early 2000s.*) |
| mono | **ft-LaBSE-MAE** | 0.62 | Elle débute dans des **séries télévisées au début des années** 1990**.** (*She made her debut in some TV Shows in the early 1990s.*) |
| reference | | | Shasta a connu une histoire explosive et éruptive. (*Shasta has had an explosive and eruptive history.*) |
| train | **LaBSE** | 0.06 | L'**histoire** de la région du Huaynaputina est marquée par un magma riche en silice**.** (*The history of the Huaynaputina region is marked by silica-rich magma.* ) |
| train | **ft-LaBSE-MAE** | 0.44 | Elle **a une histoire** très singulière**.** (*She has a very peculiar history.*) |
| mono | **LaBSE** | 0.07 | Il s'agit d'un des cônes volcaniques du mont **Shasta.** (*This is one one the vulcanic cones of Mount Shasta.*) |
| mono | **ft-LaBSE-MAE** | 0.56 | Elle **a connu une histoire** riche en rebondissements. (*Its history is full of twists and turns.*) |

**Table 11:** Two illustrations of retrieved examples for the Wikipedia test set w.r.t. the chosen cross-lingual retriever, and the retrieval pool (train set vs. monolingual corpus). Exact matches are marked in **bold**. Indicative translations of each French segment into English are also provided.

parallel data) and BT-augmented settings.[18]

| MT setting | (a) | (b) | (c) | (d) | (e) | (f) |
|---|---|---|---|---|---|---|
| **Parallel data** | | | | | | |
| train ($k = 0$) | ✓ | ✓ | - | - | - | - |
| train ($k \geq 0$) | - | - | ✓ | ✓ | ✓ | ✓ |
| test ($k \geq 0$) | - | - | ✓ | ✓ | ✓ | ✓ |
| **BT-ed data** | | | | | | |
| train ($k = 0$) | - | ✓ | - | - | - | - |
| train ($k \geq 0$) | - | - | - | ✓ | - | ✓ |
| test ($k \geq 0$) | - | - | - | - | ✓ | ✓ |

**Table 12:** Using back-translation in NMT and RANMT

Column (c) in Table 12 illustrates standard RANMT, where $k \geq 0$ parallel examples are used during training and at inference; columns (d), (e) and (f) correspond to three possible ways to integrate artificial data: (d) only during training, and retrieve selectively from the high-quality parallel data; (e) only during inference, extending a generic RANMT with relevant data; (f) both during training and at inference as in (Tezcan et al., 2024). Our use of back-translation in the experiments of Section 5 corresponds to column (e), a setting we deem representative of typical use of monolingual data in actual applications, and which crucially does not require retraining stage. Note that the same scenario is used in our CLIR approach, which, in addition, fully dispenses with the back-translation stage.

[18]We omit the case where only artificial data is used, as in fully unsupervised machine translation (Lample et al., 2018).

## H Growing the Retrieval Pool

To better analyse the effect of the retrieval pool size, we progressively increase the Wikipedia retrieval set by batches of 500K sentences from 500K to 45M, and compute the corresponding translation scores for each retriever. Figure 2 displays the evolution of the translation metrics for **LaBSE**, its fine-tuned variants and **fuzzy-bt**. Overall, for all models, the benefit of growing the retrieval pool yields an almost linear BLEU increase. COMET variations are much smaller, especially for NFA and EuroLLM where it remains nearly constant. For both metrics, **ft-LaBSE-MAE** outperforms the other setups for all data sizes in most cases. **fuzzy-bt** consistently lags behind CLIR alternatives, especially with TM$^3$-LevT.

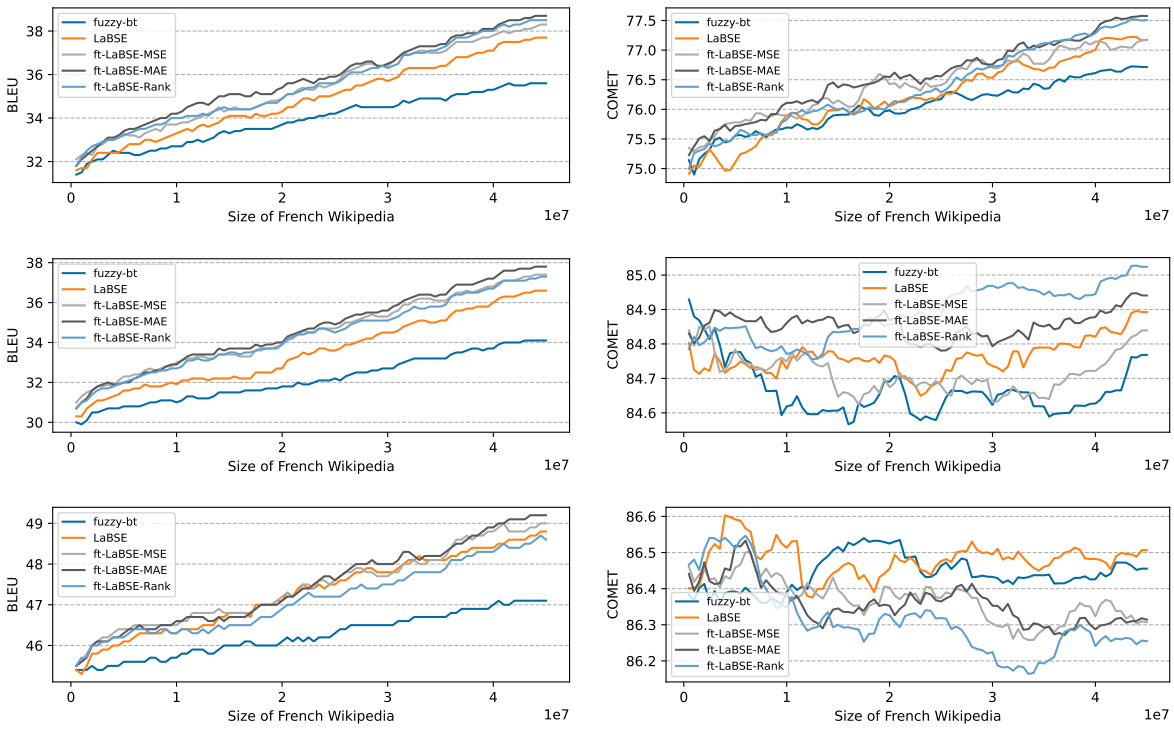


**Figure 2:** BLEU and COMET scores with a growing retrieval pool: TM$^3$-LevT (top), NFA (middle) and EuroLLM (bottom).

# I Experiments in English-Ukrainian

To further emphasize the ability of our method to retrieve relevant examples even in the absence of direct lexical overlap between source and target segments, we experiment with the translation from English into Ukrainian, with the source using the Latin script, and the target the Cyrillic script. We used the same dataset as (2024), corresponding to legal texts. The parallel corpus is split into train, validation and test sets, each containing respectively 286,417, 2000 and 1898 sentences. This corpus is augmented with a monolingual in-domain dataset of 1,461,320 Ukrainian sentences.

Our experiments are conducted with the same NFA architecture as the experiments in Section 5, as it produces translations with higher estimated-quality than the non-autoregressive system. This NFA model has been trained on 5M en–uk parallel pairs from various domains, including the in-domain train set.

We fine-tune **LaBSE** using the proposed MSE lexical loss, with examples retrieved via **fuzzy-src** from the training set. We then evaluate and compare the performance of **fuzzy-src**, **LaBSE**, and **ft-LaBSE-MSE** by computing BLEU and COMET scores of translations produced by the NFA model. For **LaBSE** and **ft-LaBSE-MSE**, we consider two retrieval settings — *train* and *train+mono* — which correspond to the datasets from which target-side examples are retrieved for the CLIR approach. Results are in Table 13, scores are computed respectively by SacreBLEU and COMET-22, significance tests also use the SacreBLEU implementations with paired bootstrap resampling (n=1000, p=0.05).

| Retrieval method | Corpus | BLEU | COMET |
|---|---|---|---|
| no example | - | 51.2 | 92.7 |
| **fuzzy-src** | train | 54.2 | 92.8 |
| **LaBSE** | train | 53.9 | 92.7 |
| **ft-LaBSE-MSE** | train | *54.3 | 92.8 |
| **LaBSE** | train+mono | 57.2 | 92.9 |
| **ft-LaBSE-MSE** | train+mono | *57.7 | 93.0 |

**Table 13:** BLEU and COMET scores obtained by the NFA model according to the retrieval method. Significance w.r.t. to the **LaBSE** counterpart is indicated by *.

We can make the following observations: (a) the baseline NFA model is about 3 BLEU points better than baseline encoder-decoder model; (b) CLIR here again matches the default RAMT setup, where retrieval only exploits the parallel training data; (c) augmenting the retrieval pool with monolingual data again induces a significant performance boost (+3.4 BLEU), when retrieval relies on the fine-tuned version of **LaBSE**. These observations are consistent

with what is reported in the main text.

## J COMET Scores

For the full picture, we also report the COMET scores of each domain for all our experiments in Tables 14, 15 and 16.

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| fuzzy-gold | 70.5 | 85.0 | 45.8 | 73.2 | 71.9 | 58.9 | 24.3 | 35.9 | 21.2 | 52.1 | 29.8 | 30.3 | -98.3 | 19.9 | 29.7 | -24.8 |
| fuzzy-src | **69.0** | **81.5** | 44.7 | **67.6** | **69.8** | 47.9 | **24.7** | **33.5** | 19.5 | 49.2 | **25.0** | 24.1 | -85.8 | 16.6 | 21.8 | -28.2 |
| fuzzy-bt | 61.6 | 75.0 | 45.6 | 55.5 | 65.8 | 37.3 | 24.1 | 30.1 | 17.7 | 45.4 | 24.2 | 11.4 | -110.6 | -11.0 | 21.3 | **-25.4** |
| LASER | 65.8 | 76.3 | 45.1 | 58.1 | 66.0 | 34.7 | 23.2 | 29.2 | **19.9** | 45.0 | 24.5 | 18.5 | -73.0 | 19.7 | **27.5** | -31.0 |
| LaBSE | 65.5 | 77.2 | 42.3 | 62.9 | 67.9 | 42.8 | 23.3 | 29.7 | 19.0 | 47.1 | 23.3 | **21.2** | -86.4 | **21.4** | 26.3 | -30.1 |
| ft-LaBSE-MSE | 68.0 | **81.1** | 44.4 | **66.5** | **69.8** | **48.0** | **24.4** | 31.9 | 18.5 | 49.6 | 23.8 | 18.5 | -76.3 | 8.7 | 15.6 | -29.1 |
| ft-LaBSE-MAE | **69.0** | 80.7 | 44.2 | 65.5 | 68.4 | 47.0 | 23.8 | **32.7** | 18.9 | **49.8** | **24.9** | 20.2 | **-70.6** | 14.1 | 19.1 | -28.6 |
| ft-LaBSE-Rank | 67.3 | 79.6 | **44.8** | 63.4 | 68.8 | 46.5 | 23.8 | 32.2 | 19.7 | 47.9 | 24.5 | 20.7 | -84.9 | 14.0 | 17.6 | -28.5 |

**Table 14:** Per domain COMET scores for TM$^3$-LevT.

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| no example | 58.9 | 55.9 | 39.9 | 48.8 | 62.3 | 42.8 | 50.1 | 45.6 | 43.9 | 43.5 | 36.4 | 82.2 | 68.9 | 83.2 | 84.7 | 80.2 |
| fuzzy-gold | 90.5 | 90.9 | 87.9 | 89.8 | 90.7 | 85.8 | 88.2 | 84.6 | 85.4 | 86.6 | 85.2 | 84.5 | 71.4 | 88.2 | 86.3 | 80.6 |
| fuzzy-src | **90.4** | **90.5** | **87.8** | **89.3** | **90.5** | 85.2 | **88.2** | **84.5** | **85.4** | 86.3 | **85.0** | **84.1** | 70.0 | **87.9** | **85.9** | **80.4** |
| fuzzy-bt | 90.0 | 90.1 | 87.7 | 88.6 | 89.9 | 84.4 | **88.2** | 84.3 | 85.2 | 86.0 | **84.9** | 83.5 | 70.1 | 87.5 | 85.2 | 80.3 |
| LASER | 90.0 | 90.1 | **87.8** | 88.8 | 90.0 | 84.4 | **88.2** | 84.2 | **85.4** | 86.4 | 84.7 | 83.5 | 70.0 | 87.4 | 85.3 | **80.4** |
| LaBSE | 90.0 | 90.3 | 87.5 | 89.0 | 89.8 | 84.7 | **88.2** | 84.1 | 85.3 | 86.4 | **84.9** | **83.8** | 70.3 | 87.2 | 85.5 | 80.3 |
| ft-LaBSE-MSE | **90.3** | **90.5** | 87.7 | **89.2** | 90.4 | 85.1 | **88.2** | **84.5** | 85.3 | **86.5** | 84.8 | **83.8** | 70.3 | **87.8** | 85.6 | **80.4** |
| ft-LaBSE-MAE | **90.3** | **90.5** | 87.7 | **89.2** | 90.4 | 85.1 | **88.2** | **84.5** | **85.4** | 86.4 | 84.8 | 83.7 | 70.2 | 87.7 | **85.7** | 80.3 |
| ft-LaBSE-Rank | 90.2 | 90.4 | 87.7 | **89.2** | **90.5** | **85.3** | **88.2** | **84.5** | 85.3 | **86.5** | **84.9** | **83.8** | **70.5** | **87.8** | **85.7** | **80.4** |

**Table 15:** Per domain COMET scores for NFA.

| | English-French | | | | | | | | | | | German-English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECB | EME | Epp | GNO | JRC | KDE | News | PHP | TED | Ubu | Wiki | KDE | Kor | JRC | EME | Sub |
| no example | 88.2 | 88.5 | 87.7 | 85.3 | 89.1 | 82.8 | 86.2 | 84.3 | 85.0 | 85.8 | 86.3 | 85.9 | 74.7 | 86.9 | 85.1 | 80.5 |
| fuzzy-gold | 90.2 | 90.9 | 87.7 | 90.1 | 90.3 | 87.5 | 86.3 | 85.2 | 85.4 | 89.0 | 87.0 | 88.0 | 75.5 | 88.4 | 87.0 | 80.8 |
| fuzzy-src | **90.0** | 90.5 | **87.8** | **89.3** | **90.1** | 86.2 | 86.2 | **85.1** | **85.3** | 88.4 | **86.8** | 86.5 | 74.9 | 87.2 | 85.5 | **80.7** |
| fuzzy-bt | 88.6 | 89.0 | **87.7** | 87.2 | 89.2 | 84.0 | 86.2 | 84.7 | **85.3** | 87.6 | 86.6 | 86.0 | 74.9 | 87.2 | 85.6 | **80.6** |
| LASER | 89.8 | 90.0 | **87.7** | 88.1 | 90.0 | 84.6 | **86.3** | 84.9 | 85.2 | 88.3 | 86.6 | 86.3 | 75.0 | **87.9** | **85.9** | **80.6** |
| LaBSE | **90.0** | 90.2 | **87.7** | 88.6 | **90.1** | 85.8 | **86.3** | 84.9 | 85.2 | 88.2 | 86.6 | 86.5 | 75.0 | **87.9** | 86.1 | 80.5 |
| ft-LaBSE-MSE | **90.0** | 90.5 | **87.7** | 88.9 | **90.1** | **86.4** | **86.3** | **85.1** | **85.3** | 88.5 | **86.8** | 86.5 | **75.1** | 87.8 | 85.8 | **80.6** |
| ft-LaBSE-MAE | 89.9 | **90.6** | **87.7** | 89.1 | **90.1** | **86.4** | 86.2 | 85.0 | 85.2 | 88.4 | 86.7 | **86.7** | **75.1** | **87.9** | 86.0 | **80.5** |
| ft-LaBSE-Rank | 89.8 | 90.3 | **87.7** | **89.2** | 90.0 | 86.2 | **86.3** | 84.9 | **85.3** | **88.6** | 86.6 | 86.6 | 75.0 | 87.8 | 86.0 | 80.5 |

**Table 16:** Per domain COMET scores for EuroLLM.