EFFICIENT HYBRID PROMPT-PRUNING FOR OPEN-SOURCE LLM BASED MACHINE TRANSLATION

Zaowad R. Abdullah [1,3], Manal Iftikhar [1,4], Md. Tariqul Islam [2,3], Rifat Shahriyar [3]

[1] Alexa Translations
[2] Verbex AI
[3] Bangladesh University of Engineering and Technology
[4] Fast Nuces Lahore

ABSTRACT

We propose a hybrid retrieval strategy for open-source LLM-based machine translation that filters out irrel- evant top-k candidates before constructing the final translation prompt, thereby reducing input token count while main- taining or improving translation quality. Throughout this work, we demonstrate that fixed top-k retrieval in translation- specific LLMs is suboptimal, often incorporating redundant or irrelevant examples into the translation prompt. Our method combines dense embedding model relevance scores and normalized sparse BM25 scores to yield a hybrid score which is later used to filter out irrelevant examples that fall below an empirically derived threshold. Unlike prior domain adaptation methods such as kNN-MT, LLM-based translation avoids dense token-level lookups. Rather, it incorporates source-translation pairs semantically/lexically similar to the translation query into the prompt and achieves a signifi- cant level of domain adaptation. While being simpler and significantly faster than kNN-MT, the quality of LLM-based MT depends highly on the context provided. Fixed retrieval configurations (e.g., top-5 or top-10), commonly adopted from general NLP tasks, often include irrelevant or redundant examples. While reranker models are usually employed to reorder retrieved examples, they still rely on a fixed top-k setup, leading to the inclusion of superfluous examples. Our experiments demonstrate a simple yet effective method that dynamically filters out suboptimal examples, retaining only the most relevant context for each translation query. Experiments across seven domains and three language pairs (DE→EN, AR→EN, ZH→EN) show that our method preserves translation performance while significantly reducing prompt size. We also compare our setup with the popular reranker model Cohere Rerank 3.5 to establish the credibility of our work. Furthermore, evaluations on the PeerQA benchmark demonstrate substantial gains in zero-shot segment- level retrieval, validating the hybrid pruning method. Our findings highlight the impact of selective example retrieval for optimally domain-adapted multilingual machine translation.

KEYWORDS

Machine Translation, LLM, RAG(Retrieval Augmented Generation), Information Retrieval,n-shot Prompt- ing, Prompt-Pruning , Domain Adaptation.

1. INTRODUCTION

Retrieval-Augmented Generation (RAG) has become a key strategy for enhancing machine translation (MT) by allowing large language models (LLMs) to incorporate external examples at infer-

David C. Wyld et al. (Eds): NATL, MLTEC, SIGEM, SEAPP, CSEA, FUZZY, DMDBS – 2025
pp. 11-27, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.152202

ence time. This paradigm is especially beneficial for domain adaptation, terminology control, and low-resource settings, where leveraging translation memories (TMs) or term bases (TBs/Glossary) improves output quality without retraining.

However, RAG-based MT's effectiveness depends on the quality of retrieved examples. Dense retrievers, such as intfloat/multilingual-e5-small [1], capture semantic similarity but of- ten miss domain-specific terms, while sparse methods like BM25 favor lexical overlap but over- look meaning. This trade-off can lead to suboptimal prompts with noisy context, inflating inference costs and reducing translation quality. Moreover, fixed-size top-k selection can introduce redun- dant or noisy context, potentially degrading translation quality and increasing inference latency. We address these issues with a lightweight, hybrid filtering strategy that selects examples based on both semantic and lexical relevance. By discarding low-quality context with a tunable threshold, our method reduces prompt size and avoids reliance on expensive rerankers.

Experiments across three language pairs (DE→EN, AR→EN, ZH→EN) and seven domains (Medical, Legal, IT, Finance, Automotive, Education and Network) show our method maintains competitive translation quality while significantly lowering memory usage and latency. The PeerQA evaluation further highlights its effectiveness in zero-shot segment retrieval, supporting the feasi- bility of scalable, retrieval-based domain adaptation using open-source LLMs.

2. RELATED WORK

Retrieval-augmented generation (RAG) has enhanced machine translation (MT), particularly in low-resource and domain-specific contexts. However, existing systems often struggle with hal- lucination and lack of generalization. In this section, we review key prior works on RAG-based translation and position our hybrid method as a practical and more efficient alternative.

2.1. THOTH AI

Miyagawa et al. [2] introduced **THOTH AI**, a RAG-based pipeline for translating Ancient Egyp- tian, an under-resourced historical language. The system integrates a vectorized lexicon and mor- phological analyzer to retrieve structured linguistic context, which is then used to guide an LLM such as Claude-3.5 Sonnet. However, the goal of this work was only to demonstrate how a care- fully crafted RAG pipeline impacts an extremely domain-specific MT setting. The effect of the reranker and superfluous context was not discussed.

2.2. RAGMT

A recent work titled **"RAG Picking Helps: Retrieval Augmented Generation for Machine Translation**" [3] introduces RAGMT, an end-to-end RAG framework that jointly trains a re- triever and a generator across translation and auxiliary tasks (Entity-Masked Language Modeling). This tightly coupled architecture improves contextual alignment but demands a large-scale train- ing setup with a complex loss function that requires vector representations of both the query and all first-stage retrieved examples. Also, jointly trained components in one domain usually don't achieve similar accuracy in another domain, which results in multiple training sessions across var- ious domains and multiple deployments of the generation model. Unlike this work, our proposed pipeline doesn't penalize superfluous context during the generation process but filters unnecessary retrievals prior to the prompt generation phase. Thus, our method only requires a domain-specific vector index and an LLM to generate high-quality translations with greater gains in BLEU-score metrics, while simultaneously avoiding the training phase, which eliminates significant computa- tional overhead. Our method is highly scalable as it needs a single generation model for all domains.

2.3.Context Overload

A third relevant study is the work titled ”**The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes**” [4] by Reimers Gurevych (2021), which examines how including too many retrieved contexts, even if semantically relevant, can negatively impact generation. The authors attribute this to information overload and the inclusion of ”hard negatives” that confuse the model. While their study proposes heuristics for managing retrieval size, it does not incorporate lexical signals into the filtering process. Our hybrid method directly addresses this by integrating BM25-based keyword relevance into the reranking and filtration step, ensuring that selected contexts are both semantically aligned and lexically coherent. Although the scope of their work was limited to demonstrating how increased index size hurts dense retrievals and favors sparse lexical retrievals, their finding motivated us to find a solution to discard superfluous context in a computationally inexpensive method.

2.4. Other Works

Our work primarily focuses on discarding noisy examples in a machine translation setup. Although the hybrid method does both reranking and filtration, finding a new reranking algorithm is not our objective. We investigated a few more works on hybrid retrieval, but none of them perform filtration. The work by **Bruch et al**. [5] describes a novel fusion method using a convex combination (CC) of lexical and semantic scores, but it requires a small set of training examples to tune its only parameter to a target domain. In a production scenario, finetuning a reranker for every possible domain is often not a viable approach. Another finding by **Rackauckas** et al. [6] demonstrates the fusion between the RAG pipeline and the RRF (Reciprocal Rank Fusion) method to produce a reciprocal score for a better final rank. This method is quite effective, but significantly time-consuming.

3. DATASETS

Our experiments span multiple domains and language pairs, utilizing both publicly available and custom-curated datasets to evaluate the effectiveness of hybrid context reduction in retrieval-augmented machine translation. We performed three types of experiments. The first setup prepares prompts with translation examples (we refer to it as TM) from RAG-DB. In the second setup, term-bases (TBs) are the prompt candidates instead of translation examples. The final setup evaluates the per- formance comparison of various methods on a standard information retrieval dataset. Below, we describe the datasets used for each experiment, including metadata and corpus statistics.

3.1. TM Experiments

For DE→EN translation tasks, we used a multi-domain dataset [7], re-split by Aharoni et al. [8]. The selected domains include Medical, Legal, and Information Technology (IT). For AR→EN, we used the Arabic Financial News dataset [9], a domain-specific corpus derived from financial news reporting. As no standard train-test split was provided, we manually created a test set and curated the retrieval corpus separately. Additionally, we conducted a separate experiment in the IT domain using a subset of the OPUS Ubuntu corpus [10].

Table 1. Dataset statistics for translation memory experiments

Domain	Language Direction	Train set	Test set
Medical	DE→EN	248099	2000
Legal	DE→EN	467309	2000
IT	DE→EN	222927	2000
Finance	AR→EN	5726	1489
Ubuntu	DE→EN	N/A	776

Corpus Statistics: The DE→EN dataset (see Table 1) consists of three specialized domains — Medical, Legal and Information Technology (IT)—each with its own training corpus and a fixed 2,000-example test set. Training set documents were used to construct the FAISS vector databases for each domain. The Ubuntu subset of Opus does not have a training set; rather, this setup reused the same FAISS index as the primary IT domain but employed a different test set consisting of 776 examples. We handcrafted the test set by removing all overlapping IT domain data (both training and testing) from Opus Ubuntu corpus. The goal was to evaluate the generalizability of the system on a strictly technical, domain-specific but different dataset. The financial dataset in Arabic comprises only 5726 training examples as RAG-DB candidates and 1489 test segments. This enables us to test our method on a low-resource setup, which is pragmatic in a production environment. The custom dataset used in this study will be released publicly upon paper acceptance.

3.2. Term Base Experiments

As shown in Table 2, for ZH→EN, we utilized structured term bases (glossaries) from the FGraDA [11] dataset to prepare the FAISS index. Unlike the DE→EN and AR→EN datasets, these experiments focus exclusively on terminology-level retrieval. This dataset covers three different domains: Automotive, Education, and Network.

Table 2. Dataset statistics for term base experiments

Domain	Language Direction	Glossary count	Test set
Automotive	ZH→EN	275	605
Education	ZH→EN	270	1309
Network	ZH→EN	360	1303

These evaluations are designed to assess how well our method supports domain-specific term alignment rather than full-sentence translation.

3.3. PeerQA

To check the generalization of our hybrid method on other tasks, we also evaluated it on **PeerQA** [12], a dataset designed to support QA and retrieval experiments in academic and scientific domains. It comprises **579** QA pairs drawn from **208** peer-reviewed articles, predominantly in Machine Learning and NLP, but also including fields such as Geoscience and Public Health. This dataset provides a strong foundation for future extensions of our work to zero-shot sentence-level information retrieval in keyword-heavy domains using hybrid retrieval techniques.

4. METHODOLOGY

In this section, we discuss the methodology in a sequential manner. It includes datastore construction, hybrid retrieval and filtering, and comparison with baselines. We explain each of steps in detail below:

4.1.Datastore Construction

For every domain analyzed in this study, we implemented vector-based retrieval indexes using the FAISS[13] library to enable efficient similarity search. The training data from each domain was encoded and stored as a domain-specific retrieval-augmented generation (RAG) datastore. In the case of the Ubuntu corpus, we reused the vector index built for the IT domain due to its semantic overlap. For the ZH→EN terminology-based experiment, we used domain glossaries as retrieval candidates instead of standard corpora.

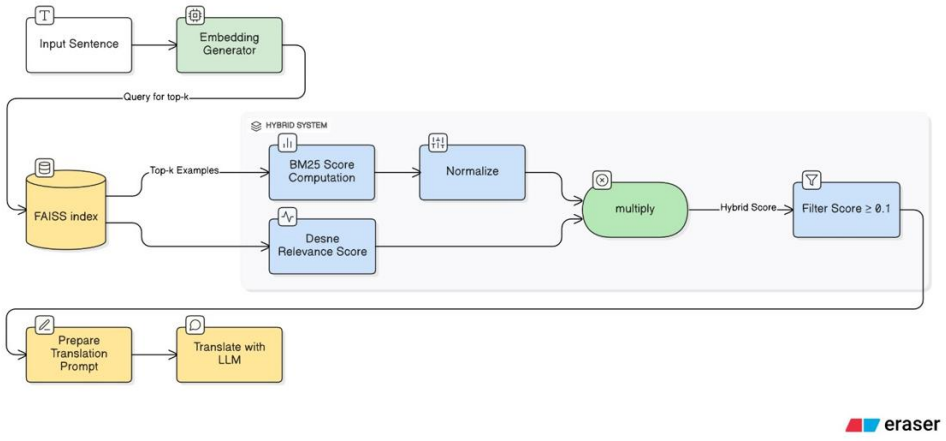

Fig. 1. Overview of our hybrid retrieval and filtering pipeline. Dense retrieval with FAISS is followed by sparse BM25 scoring and hybrid pruning before in-context translation with the LLM.

Algorithm 1: Hybrid Retrieval and Filtering for In-Context MT, as illustrated in Figure 1

Input: Query q, corpus D, top-k value k, threshold τ
Output: Filtered example set S_{filtered}

$v_q \leftarrow \text{Embed}(q)$
$E,\ s_{\text{faiss}} \leftarrow \text{FaissRetrieve}(v_q, D, k)$
$s_{\text{bm25}} \leftarrow \text{ComputeBM25Scores}(q, E)$
$s_{\min} \leftarrow \min_j s_{\text{bm25}}(q, e_j), \quad s_{\max} \leftarrow \max_j s_{\text{bm25}}(q, e_j)$
foreach $e_i \in E$ **do**
  $\hat{s}_{\text{bm25}}(q, e_i) \leftarrow \dfrac{s_{\text{bm25}}(q, e_i) - s_{\min}}{s_{\max} - s_{\min}}$
  $s_{\text{hybrid}}(q, e_i) \leftarrow \hat{s}_{\text{bm25}}(q, e_i) \cdot s_{\text{faiss}}[i]$
$S_{\text{filtered}} \leftarrow \{\, e_i \in E \mid s_{\text{hybrid}}(q, e_i) > \tau \,\}$
return S_{filtered}

4.2. Hybrid Retrieval and Filtering

At inference time, we computed dense sentence embeddings for each test query and retrieved the top-5 most similar candidates using FAISS. Each candidate's dense similarity score with the query was stored. These retrieved examples were then used to build a query-specific BM25 corpus, from which we computed keyword-based lexical relevance scores.

Since BM25 scores are unbounded and not normalized, we rescaled them to the range [0, 1]. The normalized BM25 scores were then multiplied element-wise with the corresponding FAISS similarity scores to compute a hybrid relevance score for each candidate. Candidates with hybrid scores below a threshold of 0.1 were discarded. This threshold was empirically chosen by analyzing noise reduction across 50 sampled queries from each domain.

We constructed prompts using only the retained high-quality examples and passed them to open-source LLMs for translation.

4.3. Comparison with Baselines

As detailed in the Experimental Setup section, we compare our hybrid method against multiple baselines: raw translation (no retrieval), dense similarity-based reranking, BM25-only reranking, and reranking using Cohere Rerank-3.5[14]. Except for the raw translation and our hybrid approach, all baselines construct prompts using the top-5 retrieved candidates without further filtering.

5. DEEPER DIVE INTO THE HYBRID MECHANISM AND EXAMPLES

In this section, we discuss some examples of the hybrid pruning method. This will clarify how effectively the algorithm separates semantically/lexically useful RAG candidates from irrelevant ones.

5.1. Example 1

Here the source text to translate is : ةكرشل سريفاتيملا ملاع لخدت "زناتيات لابولج Everdome Which translates into(From DeepL) : Global Titans enters Everdome's metaverse Below we attach the prompt constructed by 5 shot prompting technique. Here, the sources and targets are retrieved from the training set.

Arabic: ةعومجم LEGO حتتفت ثدحأ اهرجاتم يف ضايرلا

English: The LEGO Group Opens its Doors to More Fans across the Kingdom with the Launch of its Latest Store in Riyadh

Arabic: قوسلا يف ةلماعلا ةطاسولا تاسسؤملل لاتيباك يب يآ يج ةكرش مامضنا

English: GIB Capital Joins Saudi Stock Exchange (Tadawul) as a Member

Arabic: يمسرلا تارايسلا يعار يتينيفنإ ةمالع ليكو ةرخافلا تارايسلل ةيملاعلا تاليكوتلا ةيلودلا ةيئاسنلا ةيدوعسلا فلوجلا يتلوطبل

English: Universal Premium Motors Agencies the official dealer of INFINITI, is the Sole Automotive Sponsor of the two Saudi Ladies International Golf Tournaments

Arabic: ربع دوقولا ةرادإل سكارتلويف ةمظنأ رشنب زسيفرس روشفوأ زنايلاف ييباور ةكرش اهلوطسأ

English: Rawabi Vallianz Offshore Services is Deploying FUELTRAX Fuel Management Systems Across its Fleet

Arabic: ةكرش جاردإ "ةضباقلا سلب سنوبسير" يف رشؤم FTSE Global Micro Cap Index

English: Response Plus Holding enters the FTSE Global Micro Cap Index
Now, translate the source text below from eng to fra.

Source: ةكرشل سريفاتيملا ملاع لخدت "زناتيات لابولج" Everdome

Target:

We can see that none of the examples align closely with translation query in terms of lexical similarity. Semantic similarity is also minimum between input text and prompt candidates retrieved by embedding model. Here goes the prompt after pruning through the hybrid method. It effectively eliminated all candidates as they do not semantically/lexically align with the query.

Translate this from Arabic to English:

Arabic: ةكرشل سريفاتيملا ملاع لخدت "زناتيات لابولج" Everdome

English:

5.2. Example 2

Here the source text to translate is : ضايرلا يف ةيروفاغنسلا ةيملاعلا دلروو نو ةسردم حاتتفا

Which translates into(From DeepL) : Opening of One World International School Singapore in Riyadh

Below we attach the prompt constructed by 5 shot prompting technique.

Arabic: ضايرلا ةنيدم يف لوألا اهيّج نع نلعت نشور

English: Roshn Announces Flagship Riyadh Community

Arabic: ضايرلا يف اهل ارًقم حتتفتو ةيدوعسلا يف عسوتت اتيباكرآ

English: Arcapita Expands in Saudi Arabia, Opens Riyadh Office

Arabic: ةعومجم LEGOتفتح ثدحأ اهرجاتم يف ضايرلا

English: The LEGO Group Opens its Doors to More Fans across the Kingdom with the Launch of its Latest Store in Riyadh

Arabic: "ياك يا سا" ةعومجم عم ةكارش ناعقوت ضايرلا ةنيدمل ةيكلملا ةئيهلاو رامثتسالا ةرازو ضايرلا يف اهل ةيلود ةسردم لوأ حاتتفال ةينابسإلا

English: SEK International School to Open in Riyadh

Arabic: نأ ررقملا نمو 2022. يف ةفورعملا ةيملاعلا سرادملا ةسسؤم حتتفت ،ضايرلا ةنيدمل ةيكلملا ةئيهلا ةياعر تحت ربمتبس يف ةيدوعسلا ةيبرعلا ةكلمملا ةمصاع يف اهل يعماج مرح لوأ ةروفاغنس سرادملا بذج جمانرب راطإ يف ضايرلا يف سسؤتُ يتلا ةعبارلا ةسردملا ةيلودلا دلروو نو ةسردم نوكت نلإ ةيلودلا سرادملا باطقتسا فدهتست يتلاو ،ضايرلا ةنيدمل ةيكلملا ةئيهلا هتنبت يذلا ةيلودلا
.ةكلمملاب ميلعتلا ةرازوو رامثتسالا ةرازو عم ةكارشلاب ،ةيدوعسلا ةيبرعلا ةكلمملا ةمصاع

English: Under the patronage of Royal Commission for Riyadh City (RCRC), Singapore's prestigious Global Schools Foundation is opening its first school campus in the Saudi capital in Septem-

ber 2022. The One World International School (OWIS) will be the fourth school to be established in Riyadh under the RCRC's International Schools Attraction Program, an initiative to bring international schools to the Saudi capital in partnership with the Kingdom's Ministry of Investment and the Ministry of Education.

Now, translate the source text below from eng to fra.

Source: ضايرلا يف ةيروفاغنسلا ةيملاعلا دلروو نو ةسردم حاتتفا

Target:

Here goes the prompt constructed by the hybrid pruning method which clearly includes the opening of schools in Riyadh example and also the example with information about Singapore. It efficiently reduced prompt size by keeping the relevant examples only.

Arabic: ضايرلا يف اهل ةيلود ”ياك يإ سا” ةعومجم عم ةكارش ناعقوت ضايرلا ةنيدمل ةيكلملا ةئيهلاو رامثتسالا ةرازو ةسردم لوأ حاتتفال ةينابسإلا

English: SEK International School to Open in Riyadh

Arabic: نأ ررقملا نمو .2022 يف ةفورعملا ةيملاعلا سرادملا ةسسؤم حتتفت ،ضايرلا ةنيدمل ةيكلملا ةئيهلا ةياعر تحت ربمتبس يف ةيدوعسلا ةيبرعلا ةكلمملا ةمصاع يف اهل يعماج مرح لوأ ةروفاغنس سرادملا بذج جمانرب راطإ يف ضايرلا يف سسؤتُ يتلا ةعبارلا ةسردملا ةيلودلا دلروو نو ةسردم نوكت نلإ ةيلودلا سرادملا باطقتسا فدهتست يتلاو ،ضايرلا ةنيدمل ةيكلملا ةئيهلا هتنبت يذلا ةيلودلا
.ةكلمملاب ميلعتلا ةرازوو رامثتسالا ةرازو عم ةكارشلاب ،ةيدوعسلا ةيبرعلا ةكلمملا ةمصاع

English: Under the patronage of Royal Commission for Riyadh City (RCRC), Singapore's prestigious Global Schools Foundation is opening its first school campus in the Saudi capital in September 2022. The One World International School (OWIS) will be the fourth school to be established in Riyadh under the RCRC's International Schools Attraction Program, an initiative to bring international schools to the Saudi capital in partnership with the Kingdom's Ministry of Investment and the Ministry of Education.

Translate this from Arabic to English:

Arabic: ضايرلا يف ةيروفاغنسلا ةيملاعلا دلروو نو ةسردم حاتتفا

English:

5.3. Example 3

Here the source text to translate is : Die Verlängerung des QT-Intervalls kann Patienten dem Risiko für einen tödlichen Ausgang aussetzen.

Which translates into(From DeepL) : The prolongation of the QT interval may expose patients to the risk of a fatal outcome.

We attach the prompt constructed by 5 shot prompting technique below. Translate the following 6 examples from German to English.

Source: Eine Vorbehandlung mit Anthracyclinen kann das Risiko einer Verlängerung der QT-Zeit erhöhen.

Target: Previous treatment with anthracyclines may increase the risk of QT prolongation.

Source: Diese Ereignisse wurden vorwiegend bei Patienten mit weiteren Risikofaktoren für eine QT-Verlängerung beobachtet.

Target: These events were observed predominantly among patients with further risk factors for QTc prolongation.

Source: Besonders bei Patienten unter hohen Methadondosen sollte das Risiko einer Verlängerung der QTc-Zeit in Betracht gezogen werden.

Target: Especially in patients on a high dose of methadone, the risk for QTc prolongation should be considered.

Source: Durch QT- Verlängerung kann es zur ventrikulären Arrhythmie vom Typ Torsade de Pointes mit möglicherweise tödlichem Ausgang kommen.

Target: QT prolongation can lead to a torsade de pointes-type ventricular arrhythmia, which can be fatal.

Source: Eine Verlängerung des QT-Intervalls kann zu einem erhöhten Risiko für ventrikuläre Arrhythmien, einschließlich Torsade de pointes, führen.

Target: QT interval prolongation may lead to an increased risk of ventricular arrhythmias including Torsade de pointes.

Source: Die Verlängerung des QT-Intervalls kann Patienten dem Risiko für einen tödlichen Ausgang aussetzen.

Target:

Although all the examples are semantically close to indicate something about QT prolongation but first 3 examples don’t say anything about a fatal outcome.

Here goes the prompt constructed by our hybrid filtering method. The pruning method kept only the two pairs with semantics covering both QT prolongation condition and the outcome.

Translate the following 3 examples from German to English.

Source: Durch QT- Verlängerung kann es zur ventrikulären Arrhythmie vom Typ Torsade de Pointes mit möglicherweise tödlichem Ausgang kommen.

Target: QT prolongation can lead to a torsade de pointes-type ventricular arrhythmia, which can be fatal.

Source: Eine Verlängerung des QT-Intervalls kann zu einem erhöhten Risiko für ventrikuläre Arrhythmien, einschließlich Torsade de pointes, führen.

Target: QT interval prolongation may lead to an increased risk of ventricular arrhythmias including Torsade de pointes.

Source: Die Verlängerung des QT-Intervalls kann Patienten dem Risiko für einen tödlichen Ausgang aussetzen.

Target:

5.4. Example 4

Another solid example from the IT domain where the pruning technique dropped examples that are semantically similar but do not contain the subject “CARD” in them. Instead, they contain “Map” and “Address book”. The pruning method is efficient enough to filter out semantically similar but lexically not close enough examples.

5-shot prompt

Translate the following 6 examples from German to English.

Source: Kein Modul zur Verwaltung dieser Karte

Target: No module managing this card **Source:** Keine Karten gefunden **Target:** No maps found

Source: Keine Adressbucheinträge gefunden **Target:** No address book entries found **Source:** Keine gültige Karte

Target: no valid card

Source: Keine Karte eingeführt

Target: No card inserted

Source: Kein ATR bzw. keine Karte eingeführt

Target:

Prompt pruned by hybrid algorithm
Translate the following 4 examples from German to English.

Source: Keine gültige Karte

Target: no valid card

Source: Kein Modul zur Verwaltung dieser Karte

Target: No module managing this card **Source:** Keine Karte eingeführt **Target:** No card inserted

Source: Kein ATR bzw. keine Karte eingeführt

Target:

Only card related candidates from RAG are chosen by the hybrid method which indicates the

robustness of the algorithm.

6. EXPERIMENTAL SETUP

To evaluate the effectiveness of our proposed hybrid context reduction technique in retrieval-augmented generation (RAG) for machine translation, we designed a comprehensive experimental framework. The evaluation spans multiple domains, language pairs, retrieval strategies, and large language models (LLMs), ensuring coverage across high- and low-resource settings, diverse context types, and realistic deployment conditions.

6.1.Domains and Context Types

Our experiments cover seven specialized domains: *Medical*, *Legal*, *Information Technology (IT)*, *Finance*, *Automotive*, *Education*, and *Network*. Context types were determined based on data availability:

6.1.1. For **Medical, Legal, IT**, and **Finance**, we used full sentence-level translation memory (TM) examples, providing rich syntactic and semantic context.

6.1.2. For **Automotive, Education**, and **Network**, only terminology-level data was available. Hence, Term Base (TB) examples—consisting of short phrase-level source-target pairs—were used to maintain terminological consistency.

This design allows us to test our method under both sentence-level and terminology-level con- text construction, enabling a fair and robust evaluation across domains.

6.2. Retrieval and Reranking Techniques

Each input sentence was embedded using the `intfloat/multilingual-e5-small` model to obtain dense semantic representations. The top-5 semantically similar candidates were retrieved using a FAISS-based vector search. Following the initial retrieval stage, a secondary re-ranking phase was conducted to refine the ordering of candidate examples within the prompt according to their relevance to the input query. We evaluated three rerankers discussed below.

6.2.1. Cohere Rerank 3.5: A popular closed-source reranker model that computes semantic rele- vance scores.

6.2.2. BM25: A sparse lexical matching method.

6.2.3. Hybrid Filtering: Combines dense relevance score from FAISS and normalized BM25 scores through element-wise multiplication. A confidence threshold of `0.1` was applied to retain only high-quality examples.

6.3.Prompt Construction and Language Models

Post-filtering, the retained TM or TB examples were formatted into structured prompts and passed to one of two instruction-tuned language models, listed in Table 3:

6.3.1. TowerInstruct-13B-v1 (Unbabel) [15]: Used for DE→EN, ZH→EN, and Ubuntu IT tasks.

6.3.2. GemmaX2-28-9B-v0.1 (ModelSpace) [16]: Used for AR→EN tasks. Designed for morpho- logically rich and low-resource languages, it performs well with shorter prompts.

Table 3. LLMS used for specific language direction

Lang Dir	LLM
DE→EN	TowerInstruct-13B-v1
AR→EN	ModelSpace/GemmaX2-28-9B-v0.1
ZH→EN	TowerInstruct-13B-v1

6.4.Evaluation Metrics

We evaluated each setup using two main metrics:

6.4.1. BLEU Score: A standard metric for measuring translation quality. We used sacrebleu library to ensure reproducible scores.

6.4.2. Context Reduction Rate: The percentage of examples removed by the hybrid filtering mech- anism, measuring inference efficiency.

6.4.3. Token Reduction Rate: The percentage of tokens reduced by the hybrid filtering mechanism, measuring inference efficiency.

6.4.4. Character Reduction Rate: The percentage of character reduced by the hybrid filtering mech- anism, measuring inference efficiency.

All experiments were conducted on **NVIDIA L40 GPUs** (48 GB).

6.5. PeerQA Experiment

For the PeerQA benchmark, our objective was to test reranking effectiveness in a QA-style setting. Unlike the MT experiments, we did not apply a hard threshold to filter candidates. Instead, we reranked the top-10 retrieved examples using different strategies:

6.5.1. Raw Embedding: Top-10 based on dense similarity alone.

6.5.2. Hybrid Score: Top-10 reranked using the hybrid relevance score (Our method without the filtering phase)

6.5.3. BM25 Second-Stage: BM25 used as a second-stage reranker on dense-retrieved candidates.

6.5.4. Cohere Rerank-v3.5: Cohere used as a second-stage reranker on dense-retrieved candidates.

This evaluation helps demonstrate how our hybrid approach generalizes beyond machine trans- lation to retrieval-based generation in QA tasks.

7. RESULTS

This section presents the evaluation results for our proposed retrieval and hybrid context reduction strategies applied to machine translation across multiple language pairs and domains. Each subsection outlines the models and methods used, translation quality metrics (BLEU), context reduction statistics, and a qualitative analysis.

7.1. Translation Example(TM) Experiments Result

We evaluated the following configurations, with results reported in Table 4:

7.1.1. RAW: Baseline LLM translation without any retrieval.

7.1.2. Infloat: Embedding-based retrieval using intfloat/multilingual-e5-small.

7.1.3. Infloat + Cohere 3.5: Semantic reranking with Cohere.

7.1.4. Infloat + BM25: Reranking using BM25 scores.

7.1.5. Hybrid: Dense relevance score + normalized BM25 score fusion with threshold filtering.

Example Reduction: Initial retrieval: 5 candidates were retrieved for each query across all domains, with example reduction statistics shown in Table 5 and token and character count reduction shown in Table 6. The Queries column of Table 5 indicates test set size of each domain, Retrieved column indicates the count of 1st stage retrieval which is basically Queries*5 as we are retrieving

Table 4. BLEU Scores for TM experiments

Domain	Lang Dir	RAW	Infloat	+Cohere	+BM25	Hybrid
Medical	DE→EN	47.82	54.97	56.02	54.86	**56.40**
Legal	DE→EN	47.65	60.59	60.34	**60.69**	60.25
IT	DE→EN	39.10	44.16	43.69	43.83	**44.41**
Finance	AR→EN	27.14	45.26	45.15	45.52	**51.29**
Ubuntu	DE→EN	28.08	29.69	29.07	31.11	**35.20**

top-5 examples. Retained column indicates the number of examples remaining after hybrid pruning and finally reduction column indicates the percantage of example reducted.

In Table 6 n-tok top-5 column indicates the total number of tokens in prompts for the whole test set, n-tok hybrid column indicates the total number of tokens after hybrid pruning. Similar notion goes for n-char top 5 and n-char hybrid. The last 2 colums indicates the % redection for each of them.

Table 5. Example reduction statistics

Domain	Lang Dir	Queries	Retrieved	retained	Reduction(%)
Medical	DE→EN	2000	10000	5719	42.81
Legal	DE→EN	2000	10000	6210	37.90
IT	DE→EN	2000	10000	4654	53.46
Finance	AR→EN	1489	7445	4099	44.94
Ubuntu	DE→EN	776	3880	1187	69.49

Table 6. Token and character reduction statistics

Domain	n-tok top-5	n-tok hybrid	n-char top 5	n-char hybrid	Token reduction %	Character reduction %
Medical(DE-EN)	820272	555018	2549914	1761231	32.34	30.93
Legal(DE-EN)	1167754	853909	4176695	3068914	26.88	26.52
it(DE-EN)	434867	286099	1517683	1020918	34.21	32.73
Finance(AR-EN)	1753225	1232213	6345777	4405681	29.72	30.57

Analysis: The hybrid approach significantly improved the BLEU score over the raw baseline across most domains. The proposed method achieved the highest BLEU in the Medical, IT, Finance and Ubuntu subset, all while maintaining a significantly smaller prompt size ranging from 42.81% to 69.49% context example reduction. Although in the legal domain it was only better than the baseline score, it was very close to the other methods. But all the other methods used 10,000 initial retrievals as context. On the contrary, hybrid method used 6210 examples, a staggering 37.9% example reduction, which eventually results in significantly smaller number of input tokens, faster translation and less computational overhead. The finance dataset saw the most substantial gain, improving BLEU by almost 6 points compared to all other methods, yet using 44.94% less examples in prompt. This demonstrates its strength in low-resource (only 5726 examples in FAISS index), domain-specific translation where both lexical and semantic cues are critical. The Ubuntu corpus, being extremely tech keyword-heavy dataset, benefited significantly from aggressive context pruning. Hybrid filtering yielded a large BLEU gain while cutting over two-thirds of the RAG candidate examples.

From Table 6 we can deduce that the IT domain saw the biggest token and character count reduction of 34.21% and 32.73%. Whereas, the lowest reduction percentage is for legal domain with 26.88%(tokenwise) and 26.52%(characterwise). These results align with the example count reduction infromation. Reduced token and character count leads to less computational overhead for LLM inference and yields better throughput, all while maintaining translation quality or even surpassing the top-5 setting.

7.2.Term base (glossary) experiments. Lang Dir: ZH→EN

LLM: TowerInstruct-13B-v1

Domains: Automotive, Education, Network

Table 7. BLEU Scores for ZH→EN Term Base Translation

Domain	RAW	Hybrid
Automotive	32.87	**34.82**
Education	31.42	**32.51**
Network	19.68	**20.11**

Analysis: While gains were modest, hybrid filtering consistently outperformed the baseline in all domains, as summarized in Table 7. These results reinforce the utility of context compression even in terminology-focused scenarios. For glossary-based experiment, hybrid is the only viable method as semantic methods yield a lot of high-scoring candidates for a single TB. Sentences where there are multiple glossaries might miss example candidates in prompt with those methods because we cannot feed all available glossaries in reranked manner. Choosing hard-coded top-k means both over and under representation of glossaries. So we have to filter out unnecessary examples at the very beginning to fit within context length of LLMs. Hence the comparison is only between raw and hybrid method. we considered production scenario where feeding prefixed top-k number of examples is not an option in order to keep the prompt concise and small because a single sentence can have multiple glosseries. So we only focussed on comparing output of raw translation and hybrid method translation.

7.3.PeerQA Evaluation (Retrieval for QA)

Setting: Top-10 zero-shot retrieval (no threshold-based filtering), with results shown in Table 8:
Relative Improvement (vs. Embedding Only):

– **Hybrid:** +17.62% (NDCG), +25.09% (MAP), +8.38% (Recall), +7.32% (P@10)
– **BM25 (2-stage):** Slight gains in Recall and P@10, but lower MAP

Table 8. PeerQA Top-10 Retrieval Metrics

Method	NDCG@10	MAP@10	Recall@10	P@10
Embedding (raw)	0.0996	0.0683	0.1395	0.0321
Hybrid	**0.1172**	**0.0855**	**0.1512**	**0.0345**
BM25 (2-stage)	0.0949	0.0594	0.1456	0.0337
Cohere (2-stage)	0.0240	0.0129	0.0384	0.0110

– **Cohere:** Significant drops across all metrics

Analysis: In a zero-shot QA setting, our hybrid strategy outperformed all baselines. BM25 helped marginally but could not capture semantic alignment. Cohere underperformed, indicating that generic rerankers may not generalize well without training or tuning. This reinforces the hybrid method's adaptability across task types.

8. CONCLUSION

The first phase of results on translation quality shows us that this hybrid method contributes in two important sections. Prompt size reduction for faster inference and omitting unnecessary examples that hurt translation quality. OpenSource models are often constrained by the context length they

can offer. This method serves as a great tool to maintain smaller prompts while maintaining translation quality. The second phase of experiments on PeerQA shows that the scope of this hybrid method can be extended beyond the scope of machine translation. Wherever segment-level information retrieval is in scope and the domain is highly impacted by keywords this hybrid method can work as a great reranker as well. From prior work we are already aware about negative impact of superfluous context. This work demonstrates an efficient non-parametric method to discard unnecessary examples from first stage retrieval. Our future goal is to implement a task-agnostic filtering algorithm.

9. LIMITATION

While our method shows superior performance gain in heavily keyword-influenced domain, it might not be a good choice to use this strategy to translate generic queries. This algorithm was developed to address production scenario where the customer use case is very often formal,non-generic, and domain centric translations. For generic usecases the first stage dense retrieval works quite well most of the time. Generic reranker like Cohere comes on top when document translation is involved and placing paragraphs/full-document in proper order in prompt is crucial. Also effectiveness of this method on other tasks except translation is not thoroughly studied here. Our goal was to implement a filtering mechanism to demonstrate that translation use case is negatively impacted when there is superfluous context in keyword-heavy domain. The scope of the study is limited to segment level translation only. The same method might not be useful for other keyword-heavy task where the query is a full paragraph/document and output doesn't necessarily depend heavily on those keywords rather semantic dependency is desirable. Due to lack of domain adaptation dataset on document-level translation we could not explore hybrid method's impact on paragraph/document-level.

REFERENCES

[1] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672.

[2] So Miyagawa. 2025. RAG-enhanced neural machine translation of Ancient Egyptian text: A case study of THOTH AI. In Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 33–40, Albuquerque, USA. Association for Computational Linguistics.

[3] Anonymous. 2025. RAG picking helps: Retrieval augmented generation for machine translation. In Submitted to ACL Rolling Review December 2024. Under review.

[4] Nils Reimers and Iryna Gurevych. 2020. The curse of dense low-dimensional information retrieval for large index sizes. arXiv preprint arXiv:2012.14210.

[5] Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. ACM Trans- actions on Information Systems, 42(1):1–35.

[6] Zackary Rackauckas. 2024. Rag-fusion: A new take on retrieval augmented generation. International Journal on Natural Language Computing, 13(1):37–47.

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

[8] Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In Pro- ceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

[9] Emad A. Alghamdi, Jezia Zakraoui, and Fares A. Abanmy. 2024. Domain adaptation for arabic machine translation: Financial texts as a case study. Applied Sciences, 14(16).

[10] Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of the Eight International Confer- ence on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Associ- ation (ELRA).

[11] Wenhao Zhu, Shujian Huang, Tong Pu, Pingxuan Huang, Xu Zhang, Jian Yu, Wei Chen, Yanfeng

Wang, and Jiajun Chen. 2021. Fdmt: A benchmark dataset for fine-grained domain adaptation in machine translation. arXiv preprint arXiv:2012.15717.

[12] Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. 2025. PeerQA: A scientific question answering dataset from peer reviews. In Proceedings of the 2025 Conference of the North American Chapter of ACL (NAACL), pages 508–544, Albuquerque, NM, USA. Association for Computational Linguistics.

[13] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

[14] Cohere_Team. 2024. Introducing Rerank 3.5: Precise AI search. Cohere Blog post. Accessed 2025-07-03.

[15] Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. artins. 2024. Tower: An open multilingual large language model for translation-related tasks. Preprint, arXiv:2402.17733.

[16] Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. Preprint, arXiv:2502.02481.