Terminology-Aware Retrieval-Augmented Knowledge Distillation for Biomedical Neural Machine Translation

Maria Zafar,[1] **Souhail Bakkali,**[2] **and Rejwanul Haque**[1]
[1]Deparment of Computing, South East Technological University, Carlow, Ireland
[2]University of Rennes, CNRS, IRISA - UMR 6074, Rennes, France
{C00304209,rejwanul.haque}@setu.ie, souhail.bakkali@irisa.fr

Abstract

Knowledge distillation (KD) compresses large teacher models into smaller student models by transferring soft labels or intermediate activations. While effective in general domains, KD alone falls short in specialised machine translation (MT) settings, such as biomedical translation. The student inherits only the teacher's compressed knowledge and lacks access to external domain information. Moreover, standard KD typically relies on abundant parallel data, which is often unavailable in domain-specific scenarios. To address these limitations, we combine KD with retrieval-augmented generation (RAG) in a few-shot setting. We propose a retrieval-augmented enhanced few-shot KD framework for French-to-English biomedical translation task. The student learns to retrieve relevant in-domain knowledge from an external database, complementing the teacher's supervision. We design and compare several retrieval strategies to enhance student capacity. Experiments show that with our terminology-aware retrieval-based methods, the student achieves performance comparable to or better than the teacher, while preserving translation quality and efficiency.

1 Introduction

Large Transformer-based MT systems (Vaswani et al., 2017; Costa-Jussà et al., 2022; Devlin et al., 2019; Touvron et al., 2023) including massively multilingual encoder-decoder architectures (Xue et al., 2021; Costa-Jussà et al., 2022) and large language models (LLMs) (Touvron et al., 2023; Team et al., 2025) have shown state-of-the-art performance across diverse translation tasks. However, these models are often computationally expensive and inefficient for real-world deployment due to their substantial parameter scale and high inference latency (Pang et al., 2025; Fernandez et al., 2025). A widely adopted approach to mitigate this issue is KD (Hinton et al., 2015), which transfers knowledge from a cumbersome teacher model to a smaller and more efficient student model. Kim and Rush (2016) were the first to investigate KD techniques for NMT. Since then, a swathe of research has explored its various applications (Gu et al., 2018; Kasai et al., 2020; Gou et al., 2021; Wang et al., 2021; Zhou et al., 2021). Notably, Kim and Rush (2016) introduced an effective sequence-level KD technique for training small student models based on pseudo target sequences generated by a large teacher model. While this technique effectively decreases both the size of the model and inference time, resulting in minimum loss in performance across various benchmarks (Kasai et al., 2020; Sanh et al., 2020; Gou et al., 2021; Wang et al., 2021), standard KD techniques confine the transferred knowledge within the student's static parameters. Consequently, a shallow student model may have a limited parametric memory that cannot be easily expanded. Additionally, during inference, the student model cannot leverage the task- or domain-specific knowledge embedded in the teacher's soft labels or intermediate activations.

Retrieval-augmented generation (RAG) (Lewis et al., 2021) can address the aforementioned limitations by injecting external knowledge at in-

ference time through retrieval from a dedicated knowledge base. For instance, Zhang et al. (2023) used parametric knowledge encoded in student model's weights and non-parametric knowledge accessible via retrieval, yielding to higher-quality distilled models. More recently, Wang et al. (2025) introduced RAGtrans, a benchmark specifically designed to train and evaluate LLMs on retrieval-augmented MT using unstructured documents. Other studies have leveraged information retrieval from structured knowledge graphs (KG) to improve performance of the MT systems. For instance, Conia et al. (2024) proposed a multilingual KG-MT method to integrate information into the MT models, while Chen et al. (2024) developed a multi-agent framework that constructs internal context-based KGs in order to guide the retrieval of external information.

Translating domain-specific terms remains a major challenge in NMT, particularly in low-resource settings where terminology is critical for preserving meaning (Haque et al., 2020; Saunders, 2022; Moslem et al., 2023; Kim et al., 2024; Zheng et al., 2024). While the aforementioned retrieval-augmented methods have achieved notable success using unstructured documents or complex knowledge graphs, the potential of term-based retrieval remains largely unexplored within the KD framework. In this work, we explore the use of non-parametric memory for low-resource biomedical translation, allowing student models to leverage both teacher outputs and retrieved contextual information during inference. Specifically, we propose a retrieval-augmented enhanced few-shot KD framework. Our approach incorporates a terminology-driven non-parametric memory alongside the model's parametric weights, further refined by a test-time adaptation strategy. We conduct an extensive evaluation of this framework on the French-to-English biomedical translation task, comparing various retrieval strategies and scaling configurations. Our results demonstrate that prioritising terminology-driven retrieval significantly improves the translation of domain-specific entities, allowing compact student models to achieve performance levels comparable to much larger teacher architectures. In summary, our key contributions are as follows: (i) to address the limitations of standard distillation, we developed a retrieval-augmented KD framework for a low-resource French-to-English biomedical MT that integrates terminology-aware non-parametric memory. This grounding in an external knowledge base provides the student model with critical domain-specific context that is absent from its internal parameters, (ii) our results demonstrate that terminology-driven retrieval can effectively bridge the domain gap for shallow student models. This approach consistently outperforms standard semantic retrieval, achieving a peak average score[1] of 80.07 at $k = 15$. This represents a 0.73 point absolute gain (corresponding to a 0.92% increase) and a 3.24 point absolute gain (corresponding to a 4.22% increase) over the best-performing semantic configuration and distilled student baseline, respectively, demonstrating the framework's ability to effectively bridge the domain gap in biomedical translation, and (iii) we provide a comprehensive analysis of the optimal RAG-KD configuration, identifying that the integration of a terminology-aware external memory with k neighbours between 12 and 19 yields the most effective setup.

The remainder of the paper is organized as follows: Section 2 reviews related work in KD and retrieval-augmented translation. Section 3 details our proposed retrieval-augmented few-shot KD framework and terminology-aware retrieval strategies. Section 4 describes the experimental setup, datasets and evaluation metrics used. This section also presents our results and a comparative analysis of different retrieval configurations. Finally, Section 5 concludes the paper and discusses future research directions.

2 Related Work

There is a growing body of research focused on KD, where a compact student model is used for inference instead of a larger teacher model. Recently, Li et al. (2025) introduced LLKD, a framework designed to optimise transfer of knowledge from massive teacher models to smaller, more efficient student models through an adaptive sample selection strategy, facilitating easy deployment of high-performing student models. Generally, KD has been classified into task-specific and task-agnostic approaches. In task-specific KD, the student model is trained to replicate the teacher model's performance on a single target task (Zhang et al., 2017; Jin et al., 2019; Sun et al.,

[1]The average score refers to the mean value of multiple translation quality metrics that provides a single, consolidated measure of model performance (cf. Section 4).

2019; Mirzadeh et al., 2020; Zafar et al., 2025b). Conversely, task-agnostic KD can produce a versatile student model that may be fine-tuned for various downstream applications following the initial distillation phase (Jin et al., 2019; Sanh et al., 2020; Sun et al., 2020; Xu et al., 2020; Zhang et al., 2020; Wang et al., 2021).

Within task-specific KD, MetaDistil (Zhou et al., 2022) serves as a baseline that employs meta-learning to improve the transfer of teacher knowledge. ReAugKD (Zhang et al., 2023) extends the efficacy of task-specific KD and the fine-tuning of task-agnostic distilled models by introducing a retrieval-augmented approach. This framework transfers relational knowledge between teacher-teacher and teacher-student embedding distributions, thereby aligning the student's latent space with that of the teacher. This alignment significantly enhances the student model's ability to interface with external memory during inference. Notably, they demonstrated that ReAugKD substantially improves the student model's generalisation capabilities without the prohibitive computational overhead of retraining the entire teacher model through meta-learning.

Wang et al. (2025) focused on leveraging unstructured multilingual data, such as Wikipedia, to provide essential context for knowledge-intensive sentences. Their experimental results demonstrated significant boosts in translation quality, improving both BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) scores for English-to-Chinese and English-to-German translation tasks.

Sun et al. (2025) proposed Self-Supervised Preference Optimisation (SSPO) through which LLMs iteratively generate high-quality translation preference data for Direct Preference Optimization (DPO) (Rafailov et al., 2024) training. Their strategy focused on the iterative refinement of translation preference pairs; specifically, they demonstrated that integrating model-generated preferences with high-quality external data during DPO training yields superior performance compared to relying solely on source-side signals. SSPO's effectiveness was tested on multiple languages, domains and models, showing consistent improvements in translation performance without relying on external human or model annotations.

Ramos et al. (2025) proposed an in-context learning framework GRAMMAMT. It works by prompting an LLM with a grammatical information from Interlinear Glossed Text. Their experimental results demonstrate that GRAMMAMT yields notable improvements, especially in translation scenarios involving unseen or endangered languages.

RAG is a well-established method for tasks requiring factual accuracy and enhanced knowledge retrieval, where querying training examples during inference significantly improves likelihood. Zhan et al. (2025) presented MMRAG, a framework designed to improve the processing of LLMs for complex biomedical information. By leveraging various strategies, their systems identified the most effective reference examples to help models perform tasks like identifying drug interactions and classifying medical text. Their experiments using various retriever tools and models like Llama-3 (Touvron et al., 2023) demonstrated that structured example selection yields substantial accuracy gains over random sampling. Ultimately, their work addressed data scarcity in healthcare by refining in-context learning, emphasizing that the methodology of information retrieval and presentation is critical for high-performance results in specialised domains.

Despite the efficiency gains offered by KD, a fundamental limitation remains: the student model is restricted to the 'closed-world' knowledge of the teacher, which often lacks the specialised entities required for domain-specific tasks. To address this, recent research has moved toward augmenting the distillation process with external evidence through RAG. As discussed above, Zhang et al. (2023) complemented the teacher's supervision by providing the student with access to non-parametric memory at inference time. Conversely, while Zheng et al. (2024) emphasised the importance of terminology-enhanced prompting, it is designed for large-scale few-shot LLMs rather than the efficient, compressed student architectures required for real-world deployment. In contrast, our proposed framework introduced a terminology-aware hybrid retrieval strategy that prioritises terminological consistency – a critical requirement for maintaining semantic accuracy in specialised translation tasks. Furthermore, we moved beyond static retrieval by incorporating a dynamic test-time adaptation phase, allowing the student model to internalise query-specific domain knowledge at inference time. This unique combination ensures that our compact model achieves performance par-

ity with much larger teachers while maintaining the efficiency necessary for specialised NMT applications

3 Retrieval-Augmented Knowledge Distillation for MT

This section describes our proposed retrieval-augmented KD framework. Our framework employs a retrieval mechanism that retrieves a non-parametric memory at inference time, composed of (i) source-–target sentence pairs from the training set and (ii) translations generated by the teacher model. The MT system (student) dynamically adapts itself to each query (i.e. input sentence to be translated) at inference time by retrieving semantically similar translation examples and performing query-specific on-the-fly fine-tuning.

3.1 Semantic Indexing of Knowledge Sources

We created two separate knowledge pools and indexed them for semantic retrieval. The first pool was derived from an external monolingual corpus containing French sentences (cf. Table 1), hereafter referred to as DistilledMemory. The second pool is built from the parallel training set (cf. Table 1), referred to as ParallelMemory.

Both corpora were encoded into dense vector representations using intfloat/e5-small-v2[2] Sentence Transformer (Reimers and Gurevych, 2020).[3] This process produced fixed-dimensional embeddings that capture the semantic content of each sentence. The resulting embeddings formed two independent semantic indices, enabling efficient similarity-based retrieval during inference. The semantic indexing process is illustrated in the Initialisation Phase of Figure 1.

3.2 Retrieval at Inference Time

At inference time we encoded the query into an embedding vector again using Sentence Transformer. We then computed cosine similarity between the query embedding and every stored embedding in both knowledge pools independently. From the DistilledMemory, the top-*k* most similar source sentences are retrieved, keeping a minimum similarity threshold of 0.5 to discard weakly related examples. From the ParallelMemory, the top-*n* most similar sentences are retrieved.

[2]https://huggingface.co/intfloat/e5-small-v2

[3]https://pypi.org/project/sentence-transformers/

3.3 Few-Shot Adaptation

The retrieved exemplars served a dual purpose: providing contextual references and acting as a localised, query-specific training signal for lightweight test-time adaptation. We use a temporary deep copy of the student model without changing the original model's parameters. This temporary model copy was switched to training mode and finetuned for 10 gradient steps using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-5}. The adaptation process was conducted on the retrieved source–target pairs, which were padded and truncated to a maximum length of 256 tokens. We employed a standard cross-entropy loss – guided by the teacher's distributions – as the optimisation objective.

The dynamic adaptation allowed the student model to briefly adapt its parameters to the local semantic neighbourhood of each specific input sentence to be translated. With this, the process internalised the domain-specific vocabulary, terminology, and syntactic patterns that are characteristics of the most similar sentences in the retrieval pools. The adapted model copy existed only for the duration of a single query's translation and was discarded afterward. This dynamic few-shot adaption strategy was integrated into our inference pipeline, as illustrated in the Inference Phase of Figure 1.

3.4 Terminology-Based Few-Shot Adaptation

The standard retrieval-augmented methods often suffer from 'semantic noise', where sentences with similar syntactic structures but unrelated meanings are retrieved (Zheng et al., 2024). To ensure the adaptation process is driven by the most informationally dense segments of the input, we first extract specialised terminology from the source sentences. We used spaCy[4] to identify named entities, noun phrases and tokens tagged as proper nouns. These entities correspond to domain specific concepts such as medical entities, treatments, or substances. Furthermore, to capture clinical nuances that standard NLP terminology extractors often overlook, we applied rule-based regex patterns to identify dosages and measurement expressions (e.g., milligrams or milliliters), and temporal frequency indicators. Like dense sentence embed-

[4]A library for NLP that supports French.https://pypi.org/project/spacy/

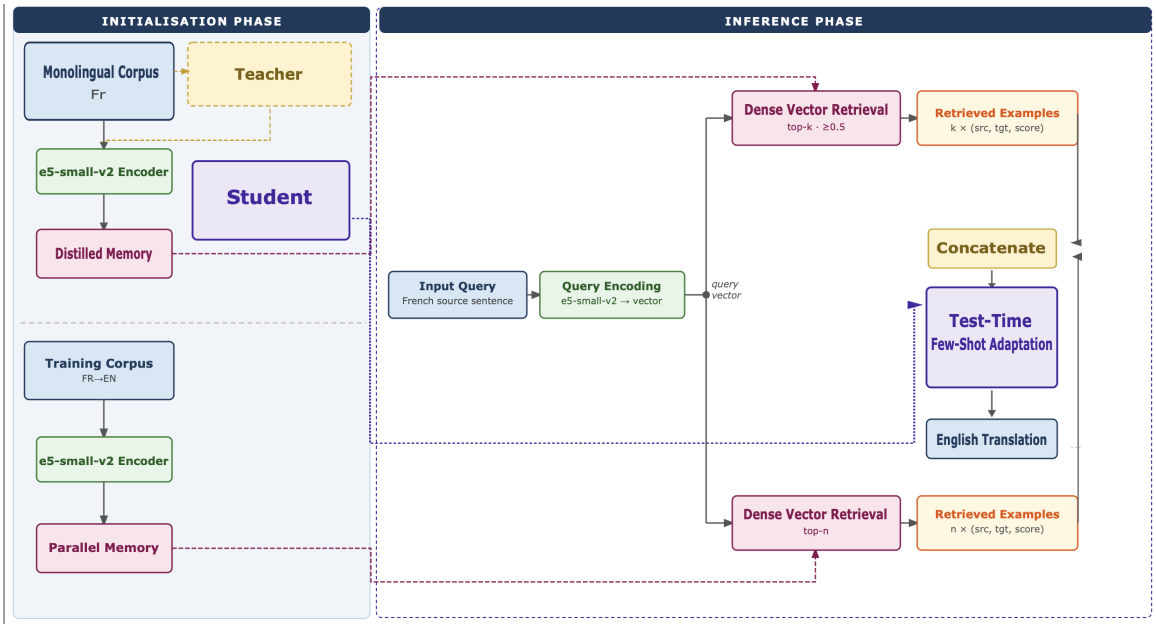

Figure 1: Architectural Overview of the Retrieval-Augmented Inference and Test-Time Adaptation Pipeline

dings, the mined terms were indexed for efficient lookup.

During inference, the query sentence undergoes the same terminology extraction and embedding computation. Retrieval is performed using a hybrid similarity function that combines terminology overlap with semantic similarity. The terminology-based score is computed using Jaccard similarity between the query's terminology set and that of each candidate sentence. The final retrieval score is defined as a weighted combination of the terminology and semantic similarities, as in (1):

$$S_{final} = \alpha \cdot Sim_{term} + (1 - \alpha) \cdot Sim_{sem} \quad (1)$$

where Sim_{term} denotes the terminology similarity, calculated as the Jaccard similarity between the sets of terms extracted from the source query and the candidate segment, Sim_{sem} represents the semantic similarity, computed as the cosine similarity between their respective sentence embeddings, and the hyperparameter $\alpha \in [0, 1]$ balances the contribution of local term-level matching against global semantic context. In our experiments, we performed a grid search over α values and observed that higher weightings for terminology similarity yielded better system performance. Specifically, we found $\alpha = 0.8$ to be the most effective, prioritising the better retrieval of domain-specific terms over broader semantic alignment.

4 Experiments

4.1 Dataset

For our experiments we used French–English (Fr–En) parallel data from the biomedical domain, ELRC-EMEA OPUS.[5] This is a medical domain data, covering a range of topics such as medicine, treatments, and pharmaceutical information. The data statistics are shown in Table 1. Prior to training, we removed duplicate sentence pairs from the corpus to ensure data quality. From the clean dataset, we sampled 1,000 and 2,000 source–target sentence pairs to create the validation and test sets, respectively. The statistics reported in Table 1 reflect the dataset after duplicate removal.

Table 1: Data Statistics

	Sentences	Vocabulary	
		French	English
Train	759,861	85,627	69,036
Valid	1,000	4,454	3,963
Test	2,000	4,557	4,052
Monolingual (Fr)	153,056	42,031	-

During inference, our MT systems (i.e., the student models) leverage non-parametric external memories implemented through a RAG framework (cf. Section 3). In addition to the parallel corpus, we used monoligual French data from ELRC-

[5]OPUS: https://opus.nlpl.eu/ELRC-EMEA/fr\&en/v1/ELRC-EMEA

Table 2: Performance of teacher, quantized teacher, and student models on the French–English biomedical test set.

Model	SacreBLEU	chrF++	COMET	BERTScore			BLEURT	Mean
				P	R	F1		
Teacher	68.33	81.40	89.23	79.82	77.27	78.49	53.28	75.40
Teacher (quantized)	70.36	83.26	90.48	83.83	81.80	82.78	55.29	78.25
Student	68.13	81.42	89.11	82.54	80.09	81.28	56.83	77.05

2720[6] and EMEA-v3.[7] We ensured that the monolingual corpus contains no overlap with the French source sentences in the parallel training data. The statistics of this data is shown in the last row of Table 1.

4.2 Baseline MT Systems

We employed MarianMT models (Junczys-Dowmunt et al., 2018) for our teacher–student training setups. As a teacher model, we used the Helsinki-NLP/opus-mt-tc-big-fr-en[8] checkpoint, which corresponds to a large Transformer model (henceforth big Transformer (BT)). For our student models, we used Helsinki-NLP/opus-mt-fr-en[9] checkpoint. From now on this model is referred as small Transformer (ST).

The training configuration is as follows: *batch size=8; eval batch size=8; learning rate=1e-5; epochs=20; save total limit=2; max sequence length=256; weight decay=applied*. We quantized the MT models using CTranslate2[10] for fast inference. The conversion was performed using the *ct2-transformers-converter* utility, which transforms MarianMT models into a runtime optimised format compatible with CTranslate2. The quantization *float16* flag was applied to reduce numerical precision from FP32 to FP16 (Rajbhandari et al., 2020).This conversion significantly reduces memory usage while improving inference speed. The best-performing quantized teacher model was then used to generate distilled data for training the student model.

We fine-tuned both BT and ST models on the training data (cf. Section 4.1) before initiating the distillation process. We then followed a standard student-teacher training setup in order to build our baseline student model, where the English translations generated by the best-performing teacher model were used as pseudo target labels for the distilled training data.

For reporting the results we used standard MT evaluation metrics: SacreBLEU (Post, 2018), chrF++ (Popović, 2015), COMET, BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). The inherent biases and variability across MT evaluation metrics are well-documented phenomena in the literature (Fomicheva and Specia, 2019; Freitag et al., 2024). Beyond standard MT evaluation metrics, we also report the average performance across all metrics to provide a composite score (Mean) for the systems. The composite score (Mean) is calculated as the arithmetic mean of the above MT evaluation metrics, as in (2):

$$\text{Mean} = \frac{S_B + C_P + C_M + B_P + B_R + B_{F1} + B_L}{7} \quad (2)$$

where S_B is SacreBLEU, C_P is chrF++, C_M is COMET, B represents BERTScore variants, and B_L is BLEURT.

The performance of the teacher model on the test set is reported in the first row of Table 2. This row corresponds to the MarianMT BT model that was obtained after fine-tuning on the domain-specific data. We also evaluated the quantized version of this model, with the corresponding results shown in the second row of Table 2. We can see from the table that the quantized model outperformed the MarianMT BT system across all evaluation metrics. We conjecture that higher performance in the quantized teacher can stem from CTranslate2 optimisation and quantization's regularization effects in technical domains (Moslem, 2025; Zafar et al., 2025a). The quantized MarianMT BT model was used to generate the distilled data for training the student model. As pointed out above, for KD, we adopted the MarianMT ST model (Junczys-Dowmunt et al., 2018) as the baseline. The third row of Table 2 reports the performance of our student model. This row refers to

[6] https://opus.nlpl.eu/datasets/ELRC-2720-EMEA?pair=fr&en$0

[7] https://opus.nlpl.eu/datasets/EMEA?pair=fr&en$0

[8] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-fr-en

[9] https://huggingface.co/Helsinki-NLP/opus-mt-fr-en

[10] CTranslate2:https://opennmt.net/CTranslate2

the MarianMT ST model that was first fine-tuned on the original training data, and then on further trained exclusively on the distilled data. When we compare the performance of the teacher and student models, we can clearly see that the performance of the student model is close to that of the teacher.

4.3 Ablation Study

We conducted our experiments using non-parametric memory (i.e. DistilledMemory) constructed from three different sources of distilled corpora (translated by the teacher model):

- DistilledMemory760K: This dataset was generated by translating the French source sentences from the original training data, comprising 759,861 sentences. Note this data was also used to train the student model.
- DistilledMemory153K: This dataset was created by translating external monolingual French sentences into English, totaling 153,056 sentences (see Table 1). This data was not used during student model training.
- DistilledMemory760+153K: This dataset is a combination of the two distilled corpora mentioned above: DistilledMemory760K and DistilledMemory153K.

These configurations allowed us to assess how different types of non-parametric memory impact the performance of the student models. Following Zhang et al. (2023), we evaluated the student models by varying the numbers of retrieved neighbours (i.e., *k* = 3, 5, 7, 9, 12, 15, 19, 23, 27, 31). As discussed in Section 3, ParallelMemory is constructed from the original training data (see Table 1). For all our experiments we utilised a single retrieved training sample (*n* = 1) from ParallelMemory as a starting point. Unlike the DistilledMemory, which contains synthetic teacher outputs, the ParallelMemory consists of gold-standard human translations. Restricting ParallelMemory to $n = 1$ prevents the compact student model from overfitting to specific neighbour patterns during the intense 10-step test-time update, thereby ensuring better domain generalisation. Exploring the impact of varying *n* is reserved for future work.

4.3.1 DistilledMemory760K

The performance of the student models while using DistilledMemory760K is reported in Table 3. Performance generally followed an upward trend with larger k values. The model achieved a chrF++ score of 81.99 at $k = 3$, eventually peaking at 83.16 when $k = 31$. COMET scores followed a similar trajectory, reaching a maximum of 90.49 at $k = 31$ The overall results indicate that the student model achieved a maximum Mean score of 78.84 when the number of retrieved neighbours (k) was set to 31.

4.3.2 DistilledMemory760K+153K

The performance of the student models while using DistilledMemory760K+153K is reported in Table 4. This configuration utilised a combined non-parametric memory source, concatenating the original training data translations with those from an external monolingual corpus. The experimental results show that performance peaked at $k = 19$ with a mean score of 78.44. This configuration provided its peak chrF++ score of 83.40 at $k = 19$. The COMET metric peaked slightly earlier at $k = 15$ with a score of 90.17. The best average performance (Mean: 78.44) was observed at $k = 19$.

4.3.3 DistilledMemory153K

The performance of the student models while using DistilledMemory153K is reported in Table 5. Note that this dataset was created by translating 153,056 external monolingual French sentences that were not used during the initial training of the student model.

Using standard semantic retrieval, the student model showed consistent gains as k increased, achieving a peak chrF++ of 84.09 at $k = 31$ and a peak COMET of 90.50 at $k = 12$. The maximum mean score was 79.34 at $k = 31$. The highest performance for this setup was observed at $k = 31$ with a mean score of 79.34.

Additionally, the performance of the student models while using DistilledMemory153K, where candidate segments were selected based on a terminology-aware composite metric is reported in Table 6. We see from the table that the performance improved significantly. This strategy yielded the highest scores, with a peak chrF++ of 84.80 and a COMET score of 90.45 at $k = 15$. This terminology-based extraction strategy outperformed all other experimental setups, reaching its highest average score of 80.07 at $k = 15$.

The experimental results show that incorporating non-parametric memory through retrieval-augmented distillation consistently improves the

Table 3: Performance of the student models on the biomedical test set while using DistilledMemory760K. Statistical significance is indicated by $*$, denoting student models that outperform the quantized teacher model with $p < 0.05$ based on bootstrap resampling of SacreBLEU scores.

k	SacreBLEU	chrF++	COMET	BERTScore			BLEURT	Mean
				P	R	F1		
3	69.13	81.99	89.94	82.38	81.22	81.76	56.44	77.55
5	69.37	82.27	89.73	82.25	80.69	81.43	55.45	77.31
7	69.99	82.81	**90.20**	83.54	81.14	82.30	54.86	77.83
9	69.81	82.25	89.84	82.52	80.70	81.56	56.16	77.54
12	70.76*	82.92	89.99	82.86	81.44	82.10	56.80	78.12
15	**71.63***	**83.31**	90.10	83.40	81.43	82.38	56.61	78.40
19	70.69*	82.99	90.10	82.65	80.98	81.78	55.80	77.85
23	71.22*	82.88	90.20	83.10	81.65	82.32	56.16	78.21
27	70.86	82.72	89.98	83.20	81.76	82.42	55.78	78.10
31	71.36	83.16	90.49	**84.17**	**82.18**	**83.13***	**57.40**	**78.84**

Table 4: Performance of student models on the biomedical test set while using DistilledMemory760K+153K. Statistical significance is indicated by $*$, denoting student models that outperform the quantized teacher model with $p < 0.05$ based on bootstrap resampling of SacreBLEU scores.

k	SacreBLEU	chrF++	COMET	BERTScore			BLEURT	Mean
				P	R	F1		
3	68.90	82.21	89.40	81.98	80.40	81.14	53.77	76.82
5	69.73	82.56	89.94	82.94	80.65	81.76	54.95	77.50
7	68.83	82.12	89.69	82.39	79.85	81.08	53.58	76.79
9	69.91	82.56	89.81	82.49	80.31	81.35	54.96	77.34
12	70.13	82.65	89.92	82.18	80.52	81.31	54.06	77.25
15	70.89	82.90	**90.17**	**83.36**	**81.52**	**82.39**	55.60	78.11
19	71.15	**83.40**	90.15	83.13	81.45	82.25	**57.60**	**78.44**
23	**71.28***	83.10	90.09	82.99	81.57	82.23	56.60	78.26
27	70.84*	83.10	90.01	83.04	81.21	82.08	55.96	78.03
31	71.12*	83.20	90.11	83.01	81.17	82.04	55.80	78.06

student model's performance across all configurations. While utilising original training data (DistilledMemory760K) and external monolingual data (DistilledMemory153K) both yielded improvements, the external dataset provided a higher peak mean score of 79.34. Most notably, the integration of a terminology-aware retrieval strategy significantly outperformed standard semantic retrieval, achieving the highest overall results with a peak chrF++ of 84.80 and a COMET score of 90.45. This strategy also reached the highest composite mean score of 80.07 at a relatively small neighbourhood size of $k = 15$, highlighting that prioritising domain-specific terminology is crucial for a narrow domain (biomedical) translation task.

When we compare the results presented in Tables 3-6, we see that DistilledMemory153K is more effective than DistilledMemory760K. The external monolingual dataset provided a higher peak mean score of 79.34 compared to 78.84. We conjecture that incorporating diverse, domain-specific external knowledge can help improve a student model's translation quality more than simply relying on its original training data.

4.3.4 Impact of Retrieval Scale (k)

Figure 2 illustrates the performance trajectories of the student models as a function of neighbourhood size (k) across six distinct plots. Each plot corresponds to a specific evaluation metric – SacreBLEU, chrF++, COMET, BERTScore (F1), and BLEURT – alongside a final plot representing the composite mean of all scores to provide a holistic assessment of the translation quality. The analysis of these curves reveals several key trends: (i) sensitivity to k: for standard semantic retrieval configurations, such as DistilledMemory760K and DistilledMemory153K, performance generally follows an upward trend as the number of retrieved neighbours increases, often peaking at the maximum tested value of $k = 31$, (ii) optimal retrieval range: the terminology-aware retrieval strategy reaches its peak performance at a much smaller retrieval scale. As shown in the plots, this method achieves its highest scores across nearly all metrics including a peak chrF++ of 84.80 and

Table 5: Performance of the student models on the biomedical test set while using DistilledMemory153K. Statistical significance is indicated by $*$, denoting student models that outperform the quantized teacher model with $p < 0.05$ based on bootstrap resampling of SacreBLEU scores.

k	SacreBLEU	chrF++	COMET	BERTScore			BLEURT	Mean
				P	R	F1		
3	71.87	83.67	90.08	83.07	82.28	82.63	58.87	78.92
5	72.38*	83.94	90.28	83.11	82.61	82.82	59.22	79.19
7	72.35*	83.92	90.35	83.17	82.34	82.71	58.88	79.10
9	72.09*	83.73	90.20	83.12	82.37	82.70	57.30	78.78
12	72.01*	83.82	**90.50**	83.33	82.41	82.82	58.21	79.01
15	71.82*	83.83	90.44	82.90	82.49	82.66	**59.58**	79.10
19	71.81*	83.85	90.43	83.18	82.47	82.79	58.69	79.03
23	71.90*	83.72	90.35	82.89	82.18	82.49	58.43	78.85
27	72.37*	83.98	90.35	83.41	82.68	83.00	58.06	79.12
31	**72.47***	**84.09**	90.43	**83.71**	**83.00**	**83.31**	58.43	**79.34**

Table 6: Performance of the student models on the biomedical test set while using DistilledMemory153K. The retrieved segments were selected based on the terminology overlap-based composite metric. Statistical significance is indicated by $*$, denoting student models that outperform the quantized teacher model with $p < 0.05$ based on bootstrap resampling of SacreBLEU scores.

Ext-k	SacreBLEU	chrF++	COMET	BERTScore			BLEURT	Mean
				P	R	F1		
3	72.93*	84.23	89.87	83.36	81.76	82.52	58.12	78.97
5	73.10*	84.45	90.09	83.47	82.29	82.84	59.12	79.34
7	**73.39***	**84.66**	90.24	83.61	82.59	83.06	59.81	79.62
9	73.03*	84.49	90.10	83.32	82.19	82.72	59.46	79.33
12	72.95*	84.66	90.35	83.96	82.75	83.32	60.28	79.75
15	73.24*	84.80	**90.45**	**84.13**	**83.12***	**83.59**	**61.15**	**80.07**
19	73.09*	84.69	90.34	83.67	82.72	83.16	60.38	79.72
23	73.13*	84.57	90.20	83.62	82.63	83.09	60.02	79.61
27	72.82*	84.47	90.21	83.47	82.61	83.00	59.38	79.42
31	72.87*	84.45	90.31	83.59	82.61	83.06	60.18	79.58

COMET of 90.45 at $k = 15$, (iii) diminishing returns: increasing k beyond the 12–19 range for the terminology-aware setup does not yield prominent improvements and, in some instances, results in a slight degradation of performance, and (iv) metric consistency: the composite Mean plot confirms that while individual metrics may peak at slightly different points (e.g., COMET peaking at $k = 12$ for standard retrieval with DistilledMemory153K), a neighbourhood size between 12 and 17 serves as an effective "sweet spot" for balancing retrieval depth with translation accuracy in terminology-heavy domains.

In order to assess the statistical significance of the SacreBLEU gains over the best-performing (quantized) teacher model, we employed bootstrap resampling (Koehn, 2004) via the standard *compare-mt*[11] evaluation toolkit. The analysis confirmed that the performance gains achieved by all student models incorporating terminology were statistically significant.

[11]Compare-MT:`https://github.com/neulab/compare-mt/tree/master`

5 Conclusion

In this work, we addressed the inherent limitations of standard KD in domain-specific NMT by proposing a terminology-aware retrieval-augmented KD framework. Our approach integrates terminology-driven non-parametric memory alongside the model's parametric weights, further refined by test-time adaption. Our experimental results on French-to-English biomedical translation demonstrate that this framework effectively bridges the domain gap for compact student models. Key findings include (a) the proposed terminology-driven retrieval consistently outperformed standard semantic-only methods, achieving a peak average score of 80.07 at $k = 15$. This represents a 3.24 point absolute gain (corresponding to a 4.22% increase) over the distilled student baseline, (b) we identified that integrating terminology-aware external memory with a neighbourhood size (k) between 12 and 19 yields the

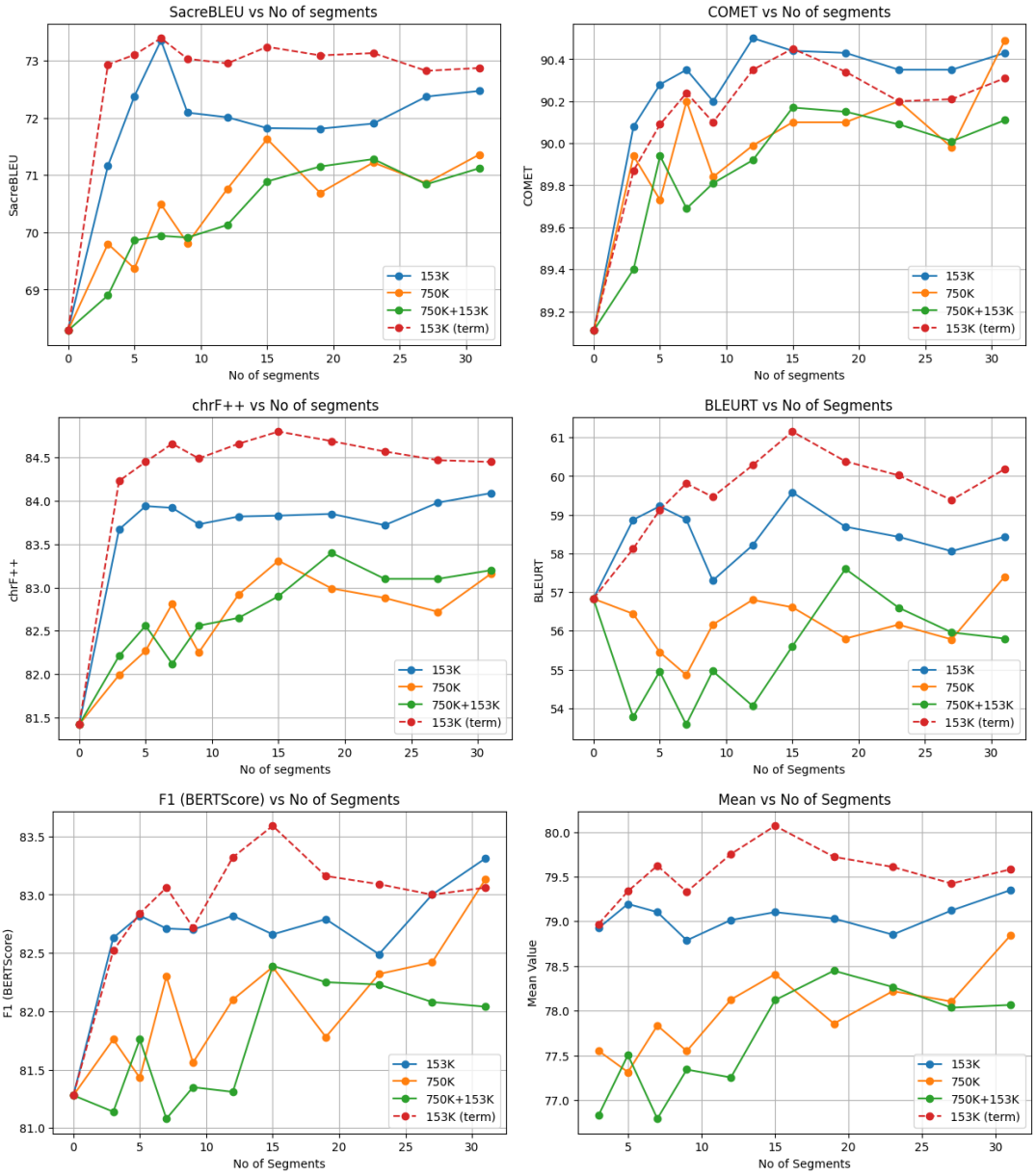

Figure 2: Effect of neighbourhood size (k) on translation quality across SacreBLEU, chrF++, COMET, BERTScore (F1), BLEURT, and their composite Mean.

most effective performance, and (c) by providing the student model with critical domain-specific context through an external knowledge base, we complemented the teacher's supervision and allowed the student to achieve performance comparable to or better than the teacher model.

In future work, we intend to evaluate the robustness of our approach by benchmarking it against a diverse suite of state-of-the-art open-weight models, including Llama 4 (Meta, 2025) and Qwen 3.5 (Yang et al., 2026). We aim to scale our experiments across a broader array of linguistic pairs and domain-specific corpora. Furthermore, we will investigate the efficacy of multi-teacher supervision by utilising a heterogeneous ensemble of LLMs to provide higher-order rationales for the student model. By using a Small Transformer student instead of a Big Transformer, we significantly reduce base inference latency. Although our test-time adaptation is streamlined using only 10 gradient steps on a minimal exemplar set ($k < 31$), we further aim to evaluate this approach through a detailed analysis of latency versus translation quality.

References

Chen, Meiqi, Fandong Meng, Yingxue Zhang, Yan Zhang, and Jie Zhou. 2024. Crat: A multi-agent framework for causality-enhanced reflective and retrieval-augmented translation with large language models.

Conia, Simone, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs.

Costa-Jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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.

Fernandez, Jared, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Luccioni, and Emma Strubell. 2025. Energy considerations of large language model inference and efficiency optimizations.

Fomicheva, Marina and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558, September.

Freitag, Markus, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, et al. 2024. Are llms breaking mt metrics? results of the wmt24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81.

Gou, Jianping, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Gu, Jiatao, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation.

Haque, Rejwanul, Mohammed Hasanuzzaman, and Andy Way. 2020. Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation*, 34(2):149–195.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Jin, Xiao, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. 2019. Knowledge distillation via route constrained optimization.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In Liu, Fei and Thamar Solorio, editors, *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Kasai, Jungo, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. 2020. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. *arXiv preprint arXiv:2006.10369*.

Kim, Yoon and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In Su, Jian, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November. Association for Computational Linguistics.

Kim, Sejoon, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Miami, Florida, USA, November. Association for Computational Linguistics.

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

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Li, Juanhui, Sreyashi Nag, Hui Liu, Xianfeng Tang, Sheikh Sarwar, Limeng Cui, Hansu Gu, Suhang Wang, Qi He, and Jiliang Tang. 2025. Learning with less: Knowledge distillation from large language models via unlabeled data.

Loshchilov, Ilya and Frank Hutter. 2019. Decoupled weight decay regularization.

Meta. 2025. Introducing llama 4: Advancing multimodal intelligence and reasoning. `https://ai.meta.com/blog/`

llama-4-multimodal-intelligence/, April. Accessed: 2026-03-26.

Mirzadeh, Seyed Iman, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198.

Moslem, Yasmin, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. *arXiv preprint arXiv:2306.01452*.

Moslem, Yasmin. 2025. Efficient speech translation through model compression and knowledge distillation. In Salesky, Elizabeth, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 379–388, Vienna, Austria (in-person and online), July. Association for Computational Linguistics.

Pang, Jianhui, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

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.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

Ramos, Rita, Everlyn Asiko Chimoto, Maartje ter Hoeve, and Natalie Schluter. 2025. Grammamt: Improving machine translation with grammar-informed in-context learning.

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.

Reimers, Nils and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Saunders, Danielle. 2022. *Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation*. Ph.D. thesis, University of Cambridge.

Sellam, Thibault, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Sun, Siqi, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression.

Sun, Zhiqing, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices.

Sun, Haoxiang, Ruize Gao, Pei Zhang, Baosong Yang, and Rui Wang. 2025. Enhancing machine translation with self-supervised preference data. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23916–23934, Vienna, Austria, July. Association for Computational Linguistics.

Team, Gemma, Aishwarya Kamath, Johan Ferret, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Fusheng, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. 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 6456–6466, Online, August. Association for Computational Linguistics.

Wang, Jiaan, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2025. Retrieval-augmented machine translation with unstructured knowledge.

Xu, Yuhui, Yuxi Li, Shuai Zhang, Wei Wen, Botao Wang, Wenrui Dai, Yingyong Qi, Yiran Chen, Weiyao Lin, and Hongkai Xiong. 2020. Trained rank pruning for efficient deep neural networks.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Yang, An, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, et al. 2026. Qwen3.5 technical report: Towards native multimodal agents and specialized reasoning. *arXiv preprint arXiv:2602.15000*.

Zafar, Maria, Souhail Bakkali, and Rejwanul Haque. 2025a. Investigating model compression for neural machine translation in the biomedical domain. In *Proceedings of the 33rd International Conference on Artificial Intelligence and Cognitive Science (AICS2025)*, Dublin, Ireland.

Zafar, Maria, Patrick J. Wall, Souhail Bakkali, and Rejwanul Haque. 2025b. Confidence-based knowledge distillation to reduce training costs and carbon footprint for low-resource neural machine translation. *Applied Sciences*, 15(14).

Zhan, Zaifu, Jun Wang, Shuang Zhou, Jiawen Deng, and Rui Zhang. 2025. Mmrag: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning.

Zhang, Ying, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2017. Deep mutual learning.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Zhang, Jianyi, Aashiq Muhamed, Aditya Anantharaman, et al. 2023. ReAugKD: Retrieval-augmented knowledge distillation for pre-trained language models. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1128–1136, Toronto, Canada, July. Association for Computational Linguistics.

Zheng, J, H Hong, F Liu, X Wang, J Su, Y Liang, and S Wu. 2024. Dragft: Adapting large language models with dictionary and retrieval augmented finetuning for domain-specific machine translation. *arXiv preprint arXiv:2402.15061v2*.

Zhou, Chunting, Graham Neubig, and Jiatao Gu. 2021. Understanding knowledge distillation in non-autoregressive machine translation.

Zhou, Wangchunshu, Canwen Xu, and Julian McAuley. 2022. Bert learns to teach: Knowledge distillation with meta learning.