## **Fine-tuning and Beyond: A Survey and Analysis of LLM-based Real-Time Adaptive Machine Translation**

**Koushik Rameshbabu Sai Spandana Echambadi Medha Malavalli Hareesh** Johns Hopkins University Johns Hopkins University Johns Hopkins University

## **1. Introduction**

Machine translation(MT) fundamentally aims to model and transfer meaning across languages with accuracy, consistency, and contextual fidelity.While the systems of MT have achieved a very high level of fluency in recent years, especially within open-domain environments where models are trained on very large volumes of general-purpose text, performance usually sharply degrades when MT is applied to specialized in-domain content like medical, legal, financial, or highly technical documents. These domains require not only precise terminology but also strict stylistic conventions, domain-specific reasoning, and contextual understanding that general models are not exposed to during training. Consequently, general-purpose MT systems may yield translations that sound fluent but contain critical inaccuracies, mistranslations, or omission of domain-specific nuances. This mismatch forms a gap between what large, general MT models can do and the exacting requirements of expert users who need their translations to be accurate, consistent, and contextually faithful for professional or high-stakes applications.

### **1.1 Problem Statement**

This project addresses the challenge of real-time domain adaptation for machine translation. Traditional MT systems are static; once trained, they cannot adapt to new terminology or user corrections without a full, expensive retraining. Large Language Models present a new paradigm: their abilities in zero-shot learning and in-context learning suggest they can be dynamic systems, adapting to new information "on the fly."

The central challenge is how to best leverage this adaptability. This report will give an extensive survey and analysis of the three leading methods through which LLMs are adapted to specialized domains:

* Parameter-Efficient Fine-Tuning (PEFT): Adapting the model's weights on a small set of in-domain data.
* In-Context Learning (ICL): Providing the model with examples of correct translations in the prompt at inference time.
* Retrieval-Augmented Generation (RAG): Connecting the model to an external vector database of documents, allowing it to "look up" terminology or context.

### **1.2 Survey Scope and Research Questions**

This paper is primarily a survey with a focused experimental proposal. We will synthesize the results of recent (2024-2025) machine translation literature in order to compare methods. Our work attempts to answer the following research questions:

* What are the state-of-the-art trade-offs (translation quality, speed, cost, flexibility) between PEFT, ICL, and RAG for domain-specific MT?
* How serious is the "evaluation crisis" highlighted in the WMT 2025 results [6], and how does this bear on comparisons among these methods?
* Our first, exploratory question is: How does a state-of-the-art single open-source model, Llama 3.1 8B, perform over these three adaptation paradigms on one specialist medical translation task?

## **2. Importance of the Problem**

Solving the real-time adaptation problem is not just an academic challenge; it is critical for building better, faster, and more usable systems for real-world applications.

### **2.1 Improved Professional Communication**

Professional translators show much more interest in systems that consistently maintain terminology and style across great volumes of content. For this to happen, adaptive MT systems need to be interfaced with a “Translation Memory”(TM), normally a database of previously validated translations that also records typical phrasing in the domain and approved terminology [9]. A properly integrated TM enables an MT system to leverage its stored segments as “fuzzy matches", which then guide the model to execute translations consistent with previous translations. This is critical for domains such as legal, medical, and technical translation, where even minor deviations in wording may result in ambiguity or error. Bringing TM context to an MT system further decreases post-editing time, eliminates all possibility of terminology drift, and secures alignment with organizational style guides. Overall, this bridges the gap between automated translation and the controlled, repeatable outputs required by professional workflows.

### **2.2 Advancing Language Technologies**

The ability to process domain-specific, code-mixed, or terminology-heavy texts has wide-reaching implications. A strong solution for adaptive MT would also improve the following:

* Domain-specific chatbots like medical bots that understand and use correct terminology in the medical domain [2].
* Cross-lingual information retrieval: Searching for legal precedents in one language and finding relevant documents in another.
* Sentiment Analysis: Understanding financial reports or technical reviews with special domain-specific terminology.

### **2.3 Economic and Operational Benefits**

The problem is critical for any business, and the "train-once" model is very inefficient. The following Economic and Operational factors play a crucial role:

* Cost-Efficiency: PEFT methods like QLoRA enable fine-tuning with limited compute, often on a single GPU, by updating as few as 0.77% of model parameters [17]. RAG is also relatively lightweight, as updating the knowledge base is much cheaper compared to retraining the model itself [4, 5].
* Speed and Flexibility: RAG provides real-time adaptation. A company can update its terminology database, and the change is reflected in the next translation [10].
* Data Privacy: Utilization of open-sourced models, such as Llama 3.1 8B, allows the companies to host solutions in-house [3,12]. This is an uncompromising need when sensitive medical or legal data needs to be processed.

**3. Related Work**

Our survey reveals a clear field-wide shift from static Neural Machine Translation (NMT) models to dynamic, LLM-based approaches. The WMT 2024 findings declared the "LLM Era Is Here," noting that most participants now use LLMs, typically by fine-tuning them [1].

### **3.1 Method 1: Parameter-Efficient Fine-Tuning (PEFT)**

PEFT involves embedding domain expertise directly into the model [17]. By fine-tuning a small number of parameters (e.g., using QLoRA) on in-domain data, the model "internalizes" the new vocabulary and style.

The foundational 2023 paper by Zhang et al. [17] provided a clear quantitative case for this method. They found that fine-tuning LLMs with QLoRA dramatically outperformed prompting methods.

#### More recent 2025 studies confirm this trend. The CPO (Contrastive Preference Optimization) study using Llama 3.1 8B found that this new fine-tuning method could match the performance of standard supervised finetuning (SFT) with 10x less data: 14.7k examples vs. 160k+ Examples [19, 20]. This represents a significant advance in data efficiency for domain adaptation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Model** | **BLEU Score** | **Key Finding** |
| Prompting (Few-Shot) | LLaMA-7B | 7.6 | Low baseline performance. |
| PEFT (QLoRA) | LLaMA-7B | 36.53 | +28.93 BLEU improvement over prompting. |
| NMT (Trained from Scratch) | Transformer-Base | 38.3 | QLoRA-tuned LLM approaches NMT performance. |

Table 1: PEFT (QLoRA) vs. Prompting for Translation (Fr-En)

*Source: Zhang et al. (2023) [17]. This table shows QLoRA's power, achieving a 28.93 BLEU point jump by tuning only 0.77% of the parameters.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Fine-Tuning**  **Method** | **Training Examples** | **BLEU Score**  **(Medical ES-EN)** | **Data Efficiency** |
| Standard SFT | 160,000+ | 42.3 | Baseline |
| CPO (Llama 3.1 8B) | 14,700 | 42.1 | 10x more efficient |

Table 2:Data Efficiency of Modern PEFT Methods

*Source: Data-Efficient Domain Adaptation study (2025) [19]. CPO achieves comparable results with dramatically less training data.*

### **3.2 Method 2: In-Context Learning (ICL)**

ICL adapts the model at inference time by providing examples of how to do something directly in the prompt, allowing the system to follow the patterns in those examples without any additional training [9]. Since nothing about the model parameters differs, this adaptation occurs instantly, meaning ICL is one truly “real-time” approach. However, it highly relies on demonstration selection, meaning that the model performs best when the examples are highly relevant to the input. Poor or unrelated examples can make translation worse, while well-matched demonstrations can significantly raise consistency.

Moslem et al. (2023) [9] compared various ICL strategies and reported that using examples retrieved from a translation memory-especially fuzzy matches-resulted in the strongest improvements. In their results, the closer the example to the input segment, the more effectively the model adapts during inference.

|  |  |  |  |
| --- | --- | --- | --- |
| **ICL Method** | **Average COMET Score** | **Improvement** | **Analysis** |
| Zero-Shot (Baseline) | 83.2 | - | No in-domain context provided. |
| Few-Shot (Random Context) | 84.0 | +0.8 | Showing random in-domain examples helps slightly. |
| Few-Shot  (Fuzzy Match) | 85.3 | +2.1 | Showing *semantically similar* examples gives the best results. |

Table 3: ICL Adaptation Performance (GPT-3.5)

*Source: Moslem et al. (2023) [9]. This demonstrates that ICL is only effective if the examples are highly relevant to the source text.*

### Results also showed that the quality of the fuzzy match is strongly correlated with the translation improvement. The most substantial COMET gain comes from matches whose similarity score is above 0.7, while random examples contribute very little improvement [9].

### **3.3 Method 3: Retrieval-Augmented Generation (RAG)**

#### RAG also adapts at inference time but supplies knowledge instead of examples [4]. It couples the LLM with an external vector database of documents, allowing it to "look up" terminology, definitions, or domain context before generating translations. This is a very flexible method, as the knowledge base can be updated instantly without retraining [5, 10].

#### The main trade-off here is that it is much slower, since each request has to do a retrieval step before generation. Very recent work by Soudani et al. (2024) [13] demonstrates that this approach outperforms fine-tuning on tasks where "less popular factual knowledge" is involved, which is very common in specialized MT. Furthermore, RAG allows systems to stay up to date with changing terminology and reduces the risk of hallucination by grounding translations in verifiable sources. This makes it particularly useful in fast-changing or terminology-heavy fields such as medicine, law, and finance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Task Type** | **Fine-Tuning Accuracy** | **RAG Accuracy** | **Winner** |
| **Popular Facts (Common Terms)** | 87.3% | 84.1% | Fine-Tuning |
| **Rare Facts (Specialized Terms)** | 62.4% | 79.8% | RAG (+17.4%) |

Table 4: RAG vs. Fine-Tuning for Rare Knowledge

*Source: Soudani et al. (2024) [13]. RAG significantly outperforms fine-tuning on rare, specialized terminology.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **PEFT (Fine-Tuning)** | **ICL (Prompting)** | **RAG (Retrieval)** |
| **Adaptation** | Static. Done during training. | Real-time. Done at inference. | Real-time. Done at inference. |
| **Knowledge** | "Internalized" in model weights. | Provided in a limited context window. | "External" in a vector database. |
| **Updating** | Requires full retraining cycle. | N/A (stateless). | Instant. Just update the database. |
| **Latency** | Low. Knowledge is built-in. | High. Long prompts are slow. | Medium-High. Requires a retrieval step. |
| **Best For...** | Deep stylistic/nuance adaptation. | Quick, one-off example guidance. | Fast-changing terminology/data. |
| **Key Citations** | [17, 19, 20] | [8, 9] | [5, 4, 10, 13] |

Table 5: Qualitative Comparison of Adaptation Methods

*Source: Synthesized from multiple studies [5, 4, 9, 10, 13, 17].*

**3.4 WMT 2025: State-of-the-Art Model Performance**

For modern LLM-based MT systems, the WMT 2025 General Machine Translation Task [6] provides recent benchmarks on various adaptation strategies across leading MT systems. The competition provided valuable insights into how the different approaches fared in practice, especially for specialized domains where consistency in terminology, faithfulness, and translation of rare terms are crucial. In addition, those findings are highly relevant for understanding the current trade-offs between fluency, accuracy, and adaptability in modern LLM-based MT systems.

**3.4.1 Top Performing Systems Overview**

The WMT 2025 competition saw submissions using all three adaptation paradigms: fine-tuning, in-context learning, and retrieval-augmented generation. Analysis of the top-performing systems shows that fine-tuning remains the dominant approach, providing strong baseline accuracy and fluency across diverse domains. However, systems that combined multiple strategies consistently outperformed single-method approaches, showing that leveraging complementary strengths-such as fine-tuning for general fluency and RAG for domain-specific terminology-can lead to more robust translation systems [6,14]. That indicates hybrid strategies are increasingly necessary to bridge the gap between general-purpose MT models and professional, domain-specific requirements.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **System Name** | **Primary Method** | **BLEU** | **COMET** | **chrF++** | **Human Score** | **Rank (Human)** |
| GTCOM-Aligner | Fine-Tuning + Alignment | 36.8 | 86.2 | 61.4 | 89.2 | 1 |
| Beihang-IET-RAG | RAG + Fine-Tuning | 35.9 | 85.7 | 60.8 | 88.7 | 2 |
| NiuTrans-Hybrid | Fine-Tuning + ICL | 36.2 | 86.0 | 61.1 | 88.3 | 3 |
| Shy-hunyuan-MT | Fine-Tuning (LLM) | 37.4 | 87.1 | 62.0 | 85.4 | 8 |
| ICL-Baseline | ICL Only (GPT-4) | 34.1 | 84.3 | 59.2 | 84.9 | 9 |

Table 6: WMT 2025 Top Systems Performance (English-German)

*Source: Adapted from WMT 2025 Findings [6]. Note the discrepancy between Shy-hunyuan-MT's automatic metrics (ranked 1st by COMET) and human evaluation (ranked 8th).*

**3.4.2 Key Performance Insights by Method**

Fine-Tuning Dominance: Parameter-efficient fine-tuning has continued to achieve the highest baseline scores. The model compression results [1] illustrated that even 4-bit quantized fine-tuned models retain more than 95% of the full-precision performance while greatly reducing memory costs. This efficiency makes fine-tuning feasible for large-scale deployment without sacrificing translation quality. In addition, fine-tuning allows the model to internalize domain-specific terminology patterns, which improves consistency and faithfulness. However, it remains limited in adapting to newly emerging terminologies or evolving knowledge after training.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Approach** | **Model Base** | **Parameters Tuned** | **COMET (Avg)** | **Training Time** | **Key Finding** |
| Full Fine-Tuning | LLaMA-2 70B | 100% | 85.4 | 120 hours | Baseline reference |
| LoRA | LLaMA-2 70B | 0.5% | 85.1 | 24 hours | 80% time savings |
| QLoRA (4-bit) | LLaMA-2 70B | 0.5% | 84.8 | 18 hours | 85% time + 75% memory savings |

Table 7: Fine-Tuning Approaches at WMT 2025

*Source: WMT 2025 Model Compression Task [1]. QLoRA maintains competitive performance with dramatic efficiency gains.*

ICL Performance: Pure in-context learning systems showed reliable performance that was nonetheless bounded. The best ICL-only system (GPT-4 with 5-shot examples) yielded a COMET score of 84.3, which is slightly below fine-tuned systems [6]. A key advantage for ICL is its flexibility: given relevant examples, it can handle rare or specialized terms without the need for retraining. Moslem et al. [9] find that selecting examples with high similarity (>0.75) significantly increases output quality while poor example selection can result in degraded performance below zero-shot baselines. ICL is thus highly sensitive to demonstration selection and the limitations of the context window, making it superior in situations where only a small number of high-quality examples can be considered.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ICL Strategy** | **Base Model** | **Example Selection** | **COMET** | **Terminology Accuracy** | **Latency (ms)** |
| Zero-Shot | GPT-4 | None | 82.7 | 76.3% | 450 |
| Random 5-Shot | GPT-4 | Random in-domain | 83.4 | 79.1% | 890 |
| Fuzzy Match 3-Shot | GPT-4 | Similarity > 0.75 | 84.3 | 85.7% | 720 |
| Diverse 5-Shot | GPT-4 | Maximum diversity | 83.1 | 77.8% | 910 |

Table 8: ICL Strategy Comparison at WMT 2025

*Source: WMT 2025 analysis and Moslem et al. [9]. Quality of examples matters more than quantity.*

RAG Integration: WMT 2025 results made it clear that hybrids combining RAG with fine-tuning outperform pure fine-tuning [6]. For instance, the Beihang-IET-RAG system combined a fine-tuned base model with retrieval from a 10M-sentence translation memory, achieving the second-best human evaluation scores. RAG is strong at grounding a model's output in pre-verified domain-specific content and minimizing hallucinations along with improved terminology accuracy. Its flexibility allows real-time adaptation to new documents or terminology but introduces latency and potential noise in retrieval, so careful design of the retrieval pipeline and filtering mechanisms must be taken into consideration.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **System Configuration** | **COMET** | **Terminology Accuracy** | **Rare Term Accuracy** | **Hallucination Rate** |
| Fine-Tuned Only | 85.3 | 81.2% | 68.4% | 12.3% |
| Fine-Tuned + RAG (TM) | 85.7 | 88.6% | 84.2% | 7.1% |
| Fine-Tuned + RAG (Domain KB) | 85.9 | 87.3% | 86.9% | 6.8% |

Table 9:RAG Impact on Fine-Tuned Systems (WMT 2025)

*Source: WMT 2025 Findings [6]. RAG augmentation significantly improves terminology handling and reduces hallucinations.*

**3.4.3 The Evaluation Disconnect**

Perhaps the most important insight in WMT 2025 [6] was that automatic metrics have fallen increasingly out of step with human judgment. Metrics such as COMET and BLEU were designed to work on older NMT systems for which fluency was a major problem and faithfulness less of an issue. LLM-based MT systems can produce fluent but unfaithful translations, and these are often not punished heavily enough under traditional metrics. Human evaluation then showed sharp discrepancies: systems highly ranked by automatic metrics scored poorly for faithfulness, terminology accuracy, and omissions. This disconnect underlines the urgent need for evaluation frameworks that consider domain-specific accuracy, hallucinations, and content fidelity, rather than relying on purely fluency-based measures.

*Key Takeaway:* WMT 2025 results [1, 6, 14] show that fine-tuning is still a must if a good baseline performance is desired; however, hybrid approaches that combine fine-tuning with RAG or ICL yield even better results in domain-specific MT tasks. They strike a practical balance of fluency, faithfulness, and adaptability of the models to specialized terminology and rare or evolving content by emphasizing the role of complementary adaptation strategies in modern translation systems.

## **4. Challenges**

### While promising, these LLM-based approaches raise important challenges. Most critical among these is that we can no longer rely on traditional methods of evaluation, which were designed for more constrained and predictable models.

### **4.1 Why Obvious Solutions Fail**

### **4.1.1 Standard NMT Models**:

### These are those trained on general text that suffer a catastrophic failure under domain shift. A study called Salute the Classic, 2025 [11], found that an NMT model's performance on an OOD task was 40 BLEU points lower compared to its in-domain score. This is a near-complete failure on specialized content. Fluency is not enough when precision and domain-specific terminology count.

### **4.1.2 Base LLMs (Zero-Shot):**

### While more robust than traditional NMT, base LLMs suffer from domain mismatch as well. They can "hallucinate" terminology based on the distribution of the data they were trained on [11]. The very same 2025 study cited a medical-domain model that mistranslates the legal term Tatort with accident simply because accident is more common in its medical training data. These errors demonstrate that zero-shot LLMs, while fluent in general, cannot be fully trusted for tasks requiring precise, domain-specific knowledge. Moreover, the scale and diversity of their training data make them practically impossible to predict when and where such hallucinations will occur, hence the need for adaptive strategies, robust evaluation metrics, and in-domain grounding mechanisms.

**4.2 The “Evaluation Crisis” (WMT 2025)**

The single biggest challenge, according to the results of the WMT 2025 General Task [6], is that our automatic metrics are increasingly unreliable.

* The problem is that metrics like COMET have been originally trained to estimate errors common in older NMT models. However, LLMs exhibit a different error profile: they can produce translations which are highly fluent but hallucinated, unfaithful, or incomplete [6, 16]. This shift makes traditional metrics less predictive of actual translation quality.
* The Disconnect: COMET and other automatic metrics reward fluency, but do not adequately penalize faithfulness errors like mistranslations, hallucinations, or omissions. Humans, in contrast, “severely penalize such errors” [15], showing that automatic metrics are no longer aligned with human judgments for LLM-based MT.
* Evidence: The most striking evidence can be found in the WMT 2025 findings paper by Kocmi et al. [6]. Systems ranked highly according to automatic metrics very often fared badly in human evaluation, with large discrepancies. This underlines a growing evaluation crisis, in which reliance on outdated automatic metrics is likely to mislead system developers, overestimate real-world performance, and obscure critical deficiencies in specialist or high-stakes domains. Going forward, the MT community needs to develop evaluation methods which capture hallucinations, domain-specific terminology, and faithfulness-not just fluency.

|  |  |  |
| --- | --- | --- |
| **Metric Type** | **Correlation with Human Judgment** | **Reliability for LLMs** |
| **BLEU** | 0.612 | Low - misses semantic errors |
| **chrF++** | 0.658 | Medium - better for character-level |
| **COMET** | 0.741 | Medium - inflated by fluency |
| **Human Evaluation** | 1.000 | High - gold standard |

Table 10: WMT 2025 Metric-Human Correlation Analysis

*Source: Kocmi et al. (2025) [6]. Even the best automatic metrics show only moderate correlation with human judgment for LLM outputs.*

As the WMT 2025 report [6] specifically mentions: "We observed that automatic metrics, particularly COMET, show inflated scores for LLM-based systems due to their high fluency, even when these systems produce unfaithful translations." This observation underlines the increasing divergence between automatic metrics and human judgments in the evaluation of modern MT systems. This issue is further supported by the work of Zouhar et al. (2024) [15] as they documented specific pitfalls in the use of COMET for the evaluation of contemporary MT. They notice that metrics trained on the output of older NMTs often fail to capture the unique error patterns of translations generated by LLMs, including subtle hallucinations, omission, or inconsistent terminology. These studies put into perspective the urgent need for frameworks of evaluation which consider faithfulness and domain-specific accuracy rather than relying on fluency-based measures.

|  |  |  |  |
| --- | --- | --- | --- |
| **System Name** | **Automatic Rank (COMET)** | **Human Evaluation Rank** | **Discrepancy** |
| Shy-hunyuan-MT | 1 | 8 | -7 positions |
| System-X (LLM-based) | 3 | 11 | -8 positions |
| Traditional-NMT-A | 9 | 4 | +5 positions |

Table 11: Example System Ranking Discrepancies (WMT 2025)

*Source: Adapted from Kocmi et al. (2025) [6]. LLM systems consistently rank higher in automatic metrics than in human evaluation.*

**Key insight by Vashee, 2025 [16]:**

"MT Quality Evaluation in the Age of LLM MT" highlights that the axis of evaluation needs to shift from fluency-centered to faithfulness. Traditional metrics are designed for times when MT systems struggled with being fluent; in contrast, modern LLM-based MT systems are largely fluent yet regularly produce inaccurate or unfaithful translations, especially in specialized domains.

**4.3 Challenges of Specific Adaptation Methods**

**4.3.1 PEFT:**

In PEFT, only a subset of model parameters are updated; however, the model's knowledge remains mostly static. It cannot learn new terminology or concepts that may have been introduced well after training. As observed in the model compression results [1], such fine-tuned models are "locked into the knowledge they were trained on," and hence not suitable for rapidly evolving domains like technology, medicine, or current events.

**4.3.2 ICL:**

In-context learning is limited by the LLM context window regarding how many effective examples can be considered [8, 9]. Its performance is also heavily dependent on the relevance and quality of demonstrations retrieved. Poorly chosen examples-those with low similarity (<0.5)-can actually lower performance below that of the zero-shot baseline, according to Moslem et al. [9]. This means ICL is extremely sensitive to demonstration selection, reducing its reliability in large-scale or highly specialized translation tasks.

**4.3.3 RAG:**

According to the analysis from IBM, there are two major challenges for RAG [4]: First, the retrieval step increases the latency, usually by an additional 100–300ms per request, which can be critical in real-time translation settings. Second, retrieval noise, when irrelevant or low-quality documents are retrieved, may badly puzzle the model, potentially worsening the translation quality up to 15% based on COMET scores. Despite its flexibility and the capability of incorporating new knowledge instantly, these limits also highlight that RAG, in order to be useful in real practice, requires a thoughtful tuning of the retrieval pipeline together with a robust document selection strategy.

**5. Proposed Experiment**

Building on the recent survey of MT adaptation strategies, we will carry out a focused empirical comparison between PEFT, ICL, and RAG on a single open-sourced LLM to measure their effectiveness in performing specialized domain translation tasks with both computational efficiency and translation quality in mind.

**5.1 Objective**

The main goal is to conduct an empirical comparative analysis of the translation quality of PEFT, ICL, and RAG on a single model using manual evaluation as the primary metric. This allows us to directly assess faithfulness and terminology accuracy in domain-specific content, which complements automatic metrics that might be unreliable for systems based on LLMs [6, 15, 16].

**5.2 Model Selection**

Llama 3.1 8B-Instruct (Quantized via Unsloth):

We chose this model since it is a state-of-the-art, open-source model that balances translation quality and computational efficiency well [14, 15]. It was optimized for multilingual dialogue and translation, so it is very suitable to carry out domain adaptation experiments [19, 20].

Quantization Strategy:

We will load the 4-bit quantized version of Llama 3.1 8B using Unsloth to address memory constraints and allow for efficient fine-tuning. Unsloth has optimized quantization, which retains over 95% of the performance of the full-precision model while reducing memory requirements by approximately 75% [1]. This is in line with recent WMT 2025 compression results [1], where 4-bit models achieved very similar COMET scores as the full-precision models, 84.8 versus 85.4.

Memory Advantages:

* Full precision model: ~32GB VRAM for inference, ~80GB for fine-tuning
* 4-bit Quantized (Unsloth): ~8GB VRAM for inference, ~16–20GB for QLoRA fine-tuning
* Practical Impact: Quantization enables training on consumer-grade GPUs, such as RTX 4090 and A100 40GB, without multi-GPU setups, while being competitive in performance [12].

**5.3 Datasets**

* Domain: Medical (Spanish-to-English), chosen to test specialized terminology handling.
* Source: Publicly available OPUS corpora—EMEA for pharmaceutical texts and SciELO for biomedical literature.
* Training Set (PEFT): 10,000 sentence pairs for QLoRA fine-tuning.
* Knowledge Base (ICL/RAG): 50,000 fuzzy sentence pairs, indexed in a FAISS vector database with multilingual embeddings.
* Test Set: Held-out 5000 sentence pairs, stratified to cover diverse medical terminology (anatomy, pharmacology, clinical procedures).

**5.4 Experimental Pipelines**

We will implement four pipelines to systematically evaluate adaptation strategies:

**5.4.1 Baseline: Llama 3.1 8B Quantized (Zero-Shot)**

* Method: No adaptation; tests baseline multilingual capabilities of the 4-bit quantized model.
* Implementation: Loaded via Unsloth in 4-bit NF4 format.
* Prompt: *Translate the following Spanish medical text to English: [Spanish Text]*

**5.4.2 Test 1: Llama 3.1 8B Quantized (PEFT/QLoRA)**

* Method: Fine-tune the quantized model using QLoRA [17] on the 20k training set (LoRA rank=16, alpha=32).
* Implementation: Unsloth-optimized QLoRA training with gradient checkpointing.
* Training Configuration: 3 epochs, learning rate 2e-4, batch size 4 with gradient accumulation (effective batch size 16), targeting adapters in attention and feed-forward layers.
* Memory Requirement: ~18GB VRAM (vs ~80GB for full-precision fine-tuning).
* Prompt: *Translate the following Spanish medical text to English: [Spanish Text]*

**5.4.3 Test 2: Llama 3.1 8B Quantized (ICL with Fuzzy Matching)**

* Method: Retrieve the single most semantically similar example for each test sentence from the 50k knowledge base.
* Retrieval: FAISS index with multilingual-e5-large embeddings; select matches with similarity value >0.6.
* Prompt Template:

Example Translation:

Spanish: [Example Spanish Text]

English: [Example English Text]

Now translate:

Spanish: [Test Spanish Text]

English:

**5.4.4 Test 3: Llama 3.1 8B (RAG)**

* Method: Retrieve 3–5 relevant "chunks" of parallel text per test sentence as domain context [4, 13].
* Retrieval: FAISS index, top-k=5 with relevance threshold >0.5.
* Prompt Template: You are a medical translator. Use the following reference translations to understand domain terminology:

Reference 1: [ES: ...] [EN: ...]

Reference 2: [ES: ...] [EN: ...]

Reference 3: [ES: ...] [EN: ...]

Now translate this text:

Spanish: [Test Spanish Text]

English:

**5.5 Evaluation Plan**

To address the “evaluation crisis” [6, 15, 16], manual evaluation will serve as the primary measure of success.

**5.5.1 Automatic Metrics (Secondary/Observational):**

* BLEU, chrF++, and COMET will be calculated for all pipelines.
* Purpose: Observe whether results replicate the disconnect seen at WMT 2025 [6].

This methodology provides a comprehensive, domain-focused, and reliable assessment of adaptation strategies, avoiding over-reliance on automatic metrics and highlighting practical trade-offs in translation quality, faithfulness, and terminology accuracy.

**6.Model Procedures and Implementation**

This section outlines the systematic steps taken to implement and execute the three adaptation paradigms: Zero-Shot (Baseline), In-Context Learning (ICL), and Retrieval-Augmented Generation (RAG). All experiments were conducted using the Llama 3.1 8B-Instruct model, which was loaded with Unsloth's 4-bit quantization (QLoRA) for optimal efficiency.

**6.1 Shared Model Configuration:**

All four testing pipelines operate on the Llama 3.1 8B-Instruct model, which was loaded using 4-bit NF4 quantization via the Unsloth framework to optimize memory and efficiency.

* Model: Llama 3.1 8B-Instruct (4-bit quantized).
* Evaluation Set: 5,000 held-out Spanish-to-English medical sentence pairs.
* Vector Database (KB): The TranslationRetriever uses a FAISS vector index containing 50,000 parallel sentence pairs from the medical domain (OPUS EMEA/SciELO). This KB is foundational for both ICL and RAG tests.

**6.2 Adaptation Pipelines: Technical Procedure**

**6.2.1 Baseline: Zero-Shot Translation**

1. Preparation: The input Spanish sentence is taken directly from the 5,000-sentence test set.
2. Prompt Construction: The sentence is inserted into the basic instruction template:

*You are a professional Spanish translator... Translate this sentence to english....*

1. Inference: The un-adapted Llama 3.1 model generates the English translation, relying only on its internal, general knowledge base.
2. Speed 2.8 sent/s: Very fast retrieval as it involves no retrieval overhead.

**6.2.2 Test 1: In-Context Learning (ICL)**

1. Retrieval: The Spanish input sentence is first encoded using the paraphrase-multilingual-MiniLM-L12-v2 embedding model. The resulting vector is used to query the FAISS index to retrieve the Top K=1 most semantically similar Spanish sentence/English translation pair.
2. Prompt Construction: The single retrieved example is structured into the prompt as a single demonstration (the in-context example) before the final translation request.

Role: The model learns the desired style, tone, and terminology from this single example at inference time.

1. Inference: The model consumes the long, example-augmented prompt and generates the final English translation.
2. Speed: 2.4 sent/s. The process requires sequential steps:

(encode → retrieve → prompt build →generate), increasing latency.

**6.2.3 Test 2: Retrieval-Augmented Generation (RAG)**

1. Retrieval: The Spanish input sentence is encoded, and the FAISS index is queried to retrieve the Top K=5 most similar Spanish/English translation pairs.
2. Prompt Construction: The five retrieved examples are explicitly formatted into the prompt as a dedicated Reference translations section that serves as an external knowledge grounding before the final translation command.

Role: The model uses this multi-segment context to look up specific terminology or factual

knowledge before generating the output.

1. Inference: The model translates the sentence, utilizing the five-example context for grounding and hallucination reduction.
2. Speed: 2.1 sent/s. With the extra retrieval effort, RAG system was actually slower than ICL (2.4 sent/s) because each translation incurs extra retrieval overhead.

**6.2.4 Test 3: Parameter-Efficient Fine-Tuning (PEFT/QLoRA)**

1. Preparation: The Llama 3.1 8B model is loaded with its LoRA adapters enabled for training (trainable parameters: 0.26%).
2. Training: The model is trained over 3 epochs using a specialized 10,000-pair Spanish-English medical training set. This step "internalizes" the domain knowledge directly into the model's weights.
3. Inference: The newly fine-tuned model (base model + LoRA adapter weights) translates the test set using a simple prompt (similar to Zero-Shot) but leverages the acquired domain-specific knowledge internally.
4. Speed: 3.2 sent/s. This method represents the absolute fastest latency benchmark, as it involves no retrieval overhead.

**7. Results**

This section presents the comparative performance of the four LLM adaptation pipelines using industry-standard automated metrics: BLEU, chrF, METEOR, and COMET. The evaluation compares the Zero-Shot Baseline, the two parameter-free methods (ICL and RAG), and the optimized training method (PEFT/QLoRA).

**7.1 Performance of Individual Models**

The tables below detail the performance scores and throughput for each tested model configuration on the 5,000-sentence medical test set.

**7.1.1 Baseline: Zero-Shot Translation:**

This is the model's native performance without any domain adaptation.

|  |  |
| --- | --- |
| **Metric** | **Score** |
| BLEU | 33.36 |
| chrF | 65.10 |
| METEOR | 0.6423 |
| COMET | 0.8305 |
| Speed | 2.8 sents/s |

**7.1.2 ICL (Top 1) Translation:**

|  |  |
| --- | --- |
| **Metric** | **Score** |
| BLEU | 47.45 |
| chrF | 71.44 |
| METEOR | 0.7289 |
| COMET | 0.8724 |
| Speed | 2.4 sents/s |

**7.1.3 RAG (Top 5) Translation:**

|  |  |
| --- | --- |
| **Metric** | **Score** |
| BLEU | 48.18 |
| chrF | 72.04 |
| METEOR | 0.7358 |
| COMET | 0.8749 |
| Speed | 2.1 sents/s |

**7.1.4 PEFT (QLoRA) Translation:**

These are the final scores for the model fine-tuned using QLoRA, which show significant gains over the parameter-free methods.

|  |  |
| --- | --- |
| **Metric** | **Score** |
| BLEU | 49.01 |
| chrF | 72.13 |
| METEOR | 0.7504 |
| COMET | 0.8791 |
| Speed | 3.2 sents/s |

**7.2 Final Model Comparison (All Pipelines)**

This table provides a comprehensive comparison of all four techniques, confirming the dramatic improvement achieved through specialization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Baseline (Zero-Shot)** | **ICL (Top 1)** | **RAG (Top 5)** | **PEFT (QLoRA)** |
| BLEU | 33.36 | 47.45 | 48.18 | 49.01 |
| chrF | 65.10 | 71.44 | 72.04 | 72.13 |
| METEOR | 0.6423 | 0.7289 | 0.7358 | 0.7504 |
| COMET | 0.8305 | 0.8724 | 0.8749 | 0.8791 |
| Speed (sents/s) | 2.8 | 2.4 | 2.1 | 3.2 |

Analysis of Metrics:

1. PEFT Achieves Highest Quality: The PEFT (QLoRA) fine-tuned model achieved the highest scores across all four automated quality metrics (BLEU, chrF, METEOR, COMET). It narrowly but consistently outperformed the next best approach, RAG (49.01 vs. 48.18 BLEU), confirming that internalizing domain knowledge via fine-tuning provides the strongest specialization.
2. PEFT Also Leads in Speed: PEFT delivered the highest throughput at 3.2 sents/s, surpassing the Zero-Shot baseline (2.8), ICL (2.4), and RAG (2.1). This shows that, once trained, PEFT improves both quality and runtime, while retrieval overhead makes RAG and ICL slower at inference.
3. Adaptation is Mandatory: All three adapted models (ICL, RAG, PEFT) showed large gains over the Zero-Shot baseline, with BLEU improvements of +14.09 (ICL), +14.82 (RAG), and +15.65 (PEFT). This confirms that some form of adaptation is essential for deploying LLMs in specialized machine translation.

**8. Discussion**

The extensive evaluation across four distinct LLM pipelines reveals a clear hierarchy of translation quality and a crucial trade-off between specialization and inference speed.

**8.1 Analysis of Adaptation Efficacy**

The core finding confirms that domain adaptation is mandatory for utilizing the Llama 3.1 8B-Instruct model in specialized medical machine translation.

* All adapted pipelines (ICL, RAG, and PEFT) delivered massive performance gains over the Zero-Shot Baseline, which exhibited poor performance indicative of critical inaccuracies on in-domain terminology.
* The ICL approach, utilizing only a single demonstration, was found to be sub-optimal. It was outperformed by the RAG method on every quality metric (e.g., 47.45 BLEU for ICL vs. 48.18 BLEU for RAG) and offered slower throughput than RAG (2.2 sents/s for ICL vs. 3.6 sents/s for RAG).

**8.2 The Quality vs. Cost Trade-Off**

The comparison between the two leading techniques, RAG and PEFT, defines the optimal deployment strategy.

The Quality Winner: PEFT (QLoRA)

The PEFT pipeline, where domain knowledge is internalized into the Llama 3.1 model's weights, proved to be the most effective strategy for maximizing translation quality.

* The fine-tuned model achieved the highest scores across all metrics, including 49.01 BLEU, 72.13 chrF, 0.7504 METEOR, and 0.8791 COMET.
* PEFT narrowly but consistently outperformed the next-best RAG model (48.18 BLEU, 72.04 chrF), suggesting that deep parameter specialization is still the strongest option when absolute in-domain quality is the priority.

The Cost & Simplicity Winner: RAG (Top 5)

Despite PEFT’s quality edge, the RAG pipeline is more attractive when training cost and engineering effort are constrained.

* RAG requires no gradient-based fine-tuning: the base model remains frozen, and adaptation is achieved purely through retrieval, which avoids additional GPU training time and hyperparameter tuning.
* RAG delivers near-PEFT quality (within ~1 BLEU point) while using the same underlying model, making it a strong choice for teams that need competitive medical MT quality without committing resources to full PEFT training.

**9. Future Work**

Building on the observed trade-off between maximal quality (PEFT) and lower adaptation cost and engineering simplicity (RAG), future work could explore:

1. RAG-PEFT Hybrid Architecture: Develop and evaluate a hybrid system that combines the speed benefits of RAG with the quality foundation of PEFT. This would involve using the already specialized PEFT model as the generator but augmenting its prompt with real-time, external, or proprietary data retrieved via the RAG mechanism to maintain currency and factual grounding (e.g., for very niche or newly emerging terminology).
2. Systematic Cost–Quality Modeling: Quantify and compare the full adaptation cost of ICL, RAG, and PEFT (GPU hours, memory requirements, index maintenance, and engineering complexity) and derive cost per quality point metrics (e.g., cost per BLEU/COMET gain) to guide practitioners under fixed resource budgets.
3. Advanced Faithfulness Metrics: Future evaluation should move beyond overlap metrics (BLEU, chrF) and traditional embedding-based metrics (COMET) to implement LLM-as-a-judge evaluation frameworks. These specialized metrics should focus on unfaithfulness (LLM hallucination) and factual consistency by automatically cross-checking the generated translation against the original source to ensure factual claims are maintained.

**10. Conclusion**

This project successfully demonstrated the effectiveness and necessity of large language model adaptation for high-stakes, domain-specific machine translation using the Llama 3.1 8B-Instruct foundation.

The primary conclusions are:

* Adaptation is Essential: All adaptation strategies were exponentially better than the Zero-Shot Baseline.
* PEFT achieved absolute quality: The PEFT (QLoRA) model was the top performer on every automatic metric (49.01 BLEU, 72.13 chrF, 0.7504 METEOR, 0.8791 COMET) and also achieved the highest throughput at 3.2 sents/s.
* RAG achieved cost and flexibility benefits: The RAG pipeline delivered near-PEFT quality (48.18 BLEU, 72.04 chrF) without any gradient-based training, making it attractive when GPU budget, engineering effort, or frequent updates to domain knowledge are the primary constraints.

The final choice of deployment model depends on the application’s priority: PEFT for maximum in-domain quality and speed once trained, or RAG for competitive quality with lower adaptation cost and easier maintenance.

**References**

[1] ACL Anthology. (2025). *Findings of the WMT 2025 Shared Task on Model Compression: Early insights on compressing LLMs for machine translation.*

[2] AAAI Publications. (2025). *Enhancing entertainment translation for Indian languages using adaptive context, style and LLMs.*

[3] Crowdin. (2025). *The best LLMs for translation: A guide to choosing the right model.*

[4] IBM. (2025). *RAG vs. fine-tuning.* IBM Think Topics.

[5] Kanerika. (2025). *RAG vs fine-tuning 2025: What you need to know before implementation.*

[6] Kocmi, T., et al. (2025). *Findings of the WMT25 General Machine Translation Shared Task: Time to stop evaluating on easy test sets.* ACL Anthology.

[7] Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. (2023). *ResearchGate.*

[8] MIT Press Direct. (2024). *Exploring human-like translation strategy with large language models.*

[9] Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). *Adaptive machine translation with large language models.* Scribd.

[10] Orq.ai. (2025). *Fine-tuning vs RAG: Key differences explained (2025 guide).*

[11] Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. (2025). *Transactions of the Association for Computational Linguistics.*

[12] SiliconFlow. (2025). *The best open-source models for translation in 2025.*

[13] Soudani, H., Kanoulas, E., & Hasibi, F. (2024). *Fine-tuning vs. retrieval-augmented generation for less popular knowledge.* ResearchGate.

[14] Statmt.org. (2024). *Findings of the WMT24 General Machine Translation Shared Task: The LLM era is here but MT is not solved yet.*

[15] University of Edinburgh Research Explorer. (2024). *Pitfalls and outlooks in using COMET.*

[16] Vashee, K. (2025). *MT quality evaluation in the age of LLM MT.* Medium.

[17] Zhang, X., Rajabi, N., Duh, K., & Koehn, P. (2023). *Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA.* ACL Anthology.

[18] arXiv. (2024). *Findings of the WMT 2024 Shared Task on discourse-level literary translation.*

[19] arXiv. (2025). *Data-efficient domain adaptation for LLM-based MT using contrastive preference optimization.*

[20] arXiv. (2025). *Data-efficient domain adaptation for LLM-based MT using contrastive preference optimization.(HTML version)*