Refining Translations with Large Language Models: A Constraint-Aware Iterative Prompting Approach

Shangfeng Chen[1], Xiayang Shi[1†], Pu Li[1], Jingjing Liu[2], Yinlin Li[3†]

[1]*School of Software Engineering, Zhengzhou University of Light Industry, No.136 Science Avenue, Zhengzhou, 450001, China*

[2]*School of Mathematics and Information Science, Zhengzhou University of Light Industry, No.5 Dongfeng Road, Zhengzhou 450003, China*

[3]*State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, No.95 Zhongguancun East Road, Beijing, 100190, China*

Abstract

Large Language Models (LLMs) have shown impressive capabilities in Machine Translation (MT), even when translating languages not specifically included in their training data. However, accurately translating rare words in low-resource or domain-specific contexts remains a significant challenge for LLMs. To address this limitation, we propose a multi-step prompt engineering approach that enhances translation accuracy by prioritizing the identification and precise rendering of critical keywords essential to semantic understanding. Our method first identifies these high-importance keywords and retrieves their translations from a bilingual dictionary, which are then integrated into the model's context via Retrieval-Augmented Generation (RAG). Additionally, we implement an iterative self-checking mechanism to mitigate potential hallucinations introduced by lengthy prompts, enabling the LLM to refine its outputs through lexical and semantic constraints. Experimental evaluations conducted using Llama and Qwen as base models on the FLORES-200 and WMT benchmarks demonstrate substantial improvements over baseline systems, particularly in low-resource settings. These results highlight the effectiveness of our approach in improving translation accuracy and consistency, offering a promising solution for enhancing MT performance in resource-constrained environments.

Keywords: Large Language Models; Machine Translation; Bilingual Dictionary; Retrieval-Augmented Generation

1. Introduction

The rapid advancement of Large Language Models (LLMs) has profoundly transformed the landscape of Natural Language Processing (NLP), particularly in the domain of Machine Translation (MT). LLMs, exemplified by models such as GPT [1], Llama [2] and BLOOM [3], possess an inherent multilingual capability that arises from extensive pre-training on diverse and expansive textual datasets. This capability enables LLMs to perform zero-shot and few-shot translation tasks, achieving impressive results even in the absence of explicit training on parallel corpora [4, 5]. Despite their promising potential, large language models (LLMs) face significant challenges in consistently

[†]Corresponding author: Xiayang Shi (Email: aryang123@163.com)
[†]Corresponding author: Yinlin Li (Email: yinlin.li@ia.ac.cn; ORCID:0000-0003-3401-1771)

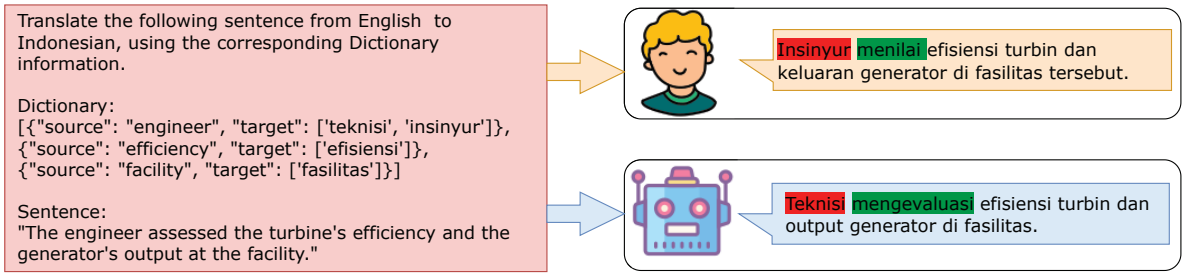


Figure 1: Comparison of human and llm translations from English to Indonesian by using dictionary information.

achieving the faithfulness and precision necessary for high-quality translation, particularly when addressing rare or specialized terminology [6, 7]. The primary limitation arises from the skewed distribution of pre-training data, which is predominantly biased toward high-resource languages. This imbalance results in an under-representation of the linguistic nuances and vocabularies associated with low-resource languages, resulting in a notable performance gap during translation tasks [8]. Furthermore, rare words often possess precise meanings within specific domains but can introduce ambiguity in varied contexts. LLMs may struggle to accurately interpret these terms due to limited contextual information, which can lead to translations that are semantically imprecise or even misleading.

The persistent limitations of LLMs in low-resource MT stem from three interconnected issues. First, the reliance on parallel corpora in traditional systems creates a data scarcity feedback loop, where insufficient data inhibits robust learning of rare linguistic structures or domain terms [9, 10]. Second, static dictionary-based approaches, which inject predefined term mappings through simplistic prompts, fail to resolve contextual polysemy. For instance, the English word “bank” may ambiguously translate to “banco” (financial) or “ribera” (river edge) in Spanish [11]. Third, current methods treat translation as a single-step process, neglecting iterative human-like revision to cross-verify terminology against contextual cues. To address these challenges, several studies have explored the integration of external knowledge into LLM-based MT pipelines to enhance translation fidelity and contextual accuracy. A prominent approach involves the incorporation of bilingual dictionaries or lexical constraints into the prompting process [12, 13]. These methods have demonstrated significant potential in improving translation precision by ensuring the consistent and accurate rendering of specialized terminology. However, they often rely on simplistic strategies, such as appending pre-translated terms to the input prompt, which neglects a more nuanced and contextually aware analysis of the source sentence. This rudimentary integration can impede the capture of the intricate semantic and syntactic nuances inherent in the original text, potentially compromising overall translation quality.

As illustrated in Figure 1, discrepancies persist between human and LLM-generated translations when employing dictionary resources. For instance, the term “engineer” maps to two Indonesian options: “teknisi” (technician) and “insinyur” (professional engineer). While the LLM selected the semantically valid “teknisi”, this choice overlooks the contextual dominance of “insinyur” in formal engineering discourse, reflecting a misalignment with domain-specific conventions. Similarly, for the term “assessed”, the model generated “mengevaluasi” despite the more colloquially

prevalent “menilai”, prioritizing literal accuracy over idiomatic fluency. These cases demonstrate that LLMs, while technically correct in handling specialized terminology, often prioritize distributional biases from training data over lexical preferences. Such mismatches arise from limitations in contextual disambiguation and an incomplete internal representation of sociolinguistic norms, ultimately reducing translation naturalness despite grammatical correctness.

Current frameworks underutilize LLMs’ potential as autonomous agents capable of iterative refinement. To address this, we propose three paradigm shifts: transitioning from passive term injection to active keyword identification using LLMs’ latent knowledge; replacing rigid dictionaries with context-aware Retrieval-Augmented Generation (RAG) [14] for dynamic term alignment; and introducing dual-agent refinement, where LLMs alternate between translator and editor roles to emulate professional human workflows. Through comprehensive experiments on the FLORES-200 [15] benchmark dataset for low-resource languages and contamination-free WMT datasets [16, 17, 18], we demonstrate that our method significantly outperforms existing approaches, achieving state-of-the-art results across multiple language pairs. Our main contributions are three-fold:

- **Keyword Identification and Constraint Adherence**: The proposed method effectively identifies keywords crucial to translation quality while filtering out less relevant terms, ensuring adherence to lexical constraints.
- **Enhanced Retrieval-Augmented Translation**: By integrating RAG to incorporate bilingual dictionary translations into the LLM’s context window and applying a post-translation self-checking mechanism, the approach minimizes misunderstandings and optimizes translation accuracy and adherence to lexical constraints.
- **Prompt-Based Iterative Refinement**: Leveraging prompt-based techniques, our approach advances MT tasks with LLMs without the need for fine-tuning, resulting in notable improvements in translation performance.

2. Related Work

2.1. Prompt of LLM for MT

Prompting techniques have gained significant attention as a powerful strategy for leveraging the capabilities of large language models (LLMs) in machine translation (MT). Early research demonstrated the efficacy of LLMs in executing zero-shot and few-shot translation tasks using simple prompts, even in the absence of parallel corpora [19, 20, 21]. Building on this, Vilar [22] advanced the field by using prompts to control various translation attributes, such as formality and dialect. Further research has focused on the effectiveness of few-shot learning paradigms in MT, with particular emphasis on the critical role of prompt engineering in achieving high-quality translations, especially for low-resource languages [23]. Recent studies have also explored the strategic selection of in-context examples to enhance translation quality [24, 25]. By choosing examples that align closely with the target translation context, these works demonstrated that LLMs can generate more accurate and contextually appropriate translations, even in specialized domains. In parallel, other studies have explored the incorporation of external knowledge such as knowledge graphs to improve machine translation performance [26, 27]. These approaches show that enhancing LLMs

with structured, domain-specific knowledge allows for more precise translations, especially in cases involving technical terms and specialized contexts. In the realm of knowledge-intensive tasks, the Retrieval-Augmented Generation (RAG) [14] framework integrates external knowledge retrieval into the generation process, allowing LLMs to incorporate relevant information from external databases or documents and produce more accurate, contextually appropriate responses. Moreover, the RAMP [28] framework significantly improves attribute-controlled translation by amalgamating retrieval-augmented prompts with LLMs, thereby facilitating the generation of more contextually accurate and nuanced translations.

2.2. Lexical-based in MT

Lexical augmentation through bilingual dictionaries remains pivotal in addressing rare-term translation challenges, particularly for low-resource MT scenarios. Early work by Zhang [29] established a paradigm for neural machine translation enhancement through systematic dictionary integration, demonstrating its efficacy in bridging lexical gaps caused by training data sparsity. Subsequent studies extended this principle across multiple dimensions: Unsupervised MT frameworks [30, 31] leveraged dictionary-derived phrase pairs as anchor points for cross-lingual alignment, proving critical in zero-resource settings; Data synthesis techniques [32] utilized dictionaries to generate pseudo-parallel corpora, enhancing model generalization despite limited authentic data; Pre-training integration exemplified by Lin's [33] Bilingual Dictionary-based Language Model (BDLM), which encodes translation pairs directly into latent representations through contrastive dictionary learning. While these methods substantially enhance term-level accuracy, two critical limitations remain unresolved [34, 35]. Firstly, static dictionary injection, whether achieved through prompt engineering or data augmentation, fails to address contextual polysemy, often necessitating manual disambiguation [36]. Secondly, the integration of dictionaries and prompts may lead to overly lengthy contextual information, which can cause LLMs to generate non-target language hallucinations, particularly when processing morphologically rich languages with complex syntactic structures.

2.3. Reflection on LLM

Reflection mechanisms in LLMs have emerged as a pivotal area of research, aiming to enhance the reliability and accuracy of generated outputs by enabling models to introspect, verify, and refine their responses. Wang [37] introduced a self-consistency technique that generates multiple potential solutions for a given problem and selects the most consistent one, thereby improving reasoning tasks through ensemble-like methods. Yao [38] proposed the ReAct framework, which integrates reasoning with actionable steps, enabling LLMs to adjust responses dynamically based on contextual feedback and enhancing their adaptability. Additionally, Lightman [39] explored a self-verification mechanism where each reasoning step is iteratively cross-checked against external knowledge sources, significantly boosting factual accuracy in multi-hop reasoning tasks. In the context of fact-checking, Lee [40] embedded self-verification capabilities within LLMs, allowing real-time assessment and correction of factual inaccuracies, which are crucial for applications such as news generation and academic writing. Additionally, the Reflexion framework [41]

leverages linguistic feedback rather than direct model updates to facilitate continuous improvement, demonstrating substantial performance gains across various tasks and highlighting the potential of feedback-driven refinement in enhancing model capabilities and adaptability.

3. Methodology

This section outlines the core principles and implementation details of our proposed methodology, providing a clear overview of the technique and the rationale behind its design.

3.1. Overview of Proposed Method

Our methodology introduces a three-stage framework to enhance translation fidelity through contextually grounded lexical alignment, as illustrated in Figure 2. In step 1, leveraging the latent linguistic knowledge of LLMs, we dynamically identify keywords within the source text by evaluating semantic saliency scores proportional to sentence length and term specificity. In step 2, prioritized keywords are encoded into dense vector representations to retrieve contextually appropriate translations from bilingual dictionaries via similarity-based retrieval. These aligned term pairs are then synthesized with the source sentence within structured prompts, guiding LLMs to generate initial translations. In step 3, a dual-agent refinement process emulates professional translation workflows: the LLM alternates between generating current translations and iteratively editing outputs as an autonomous proofreader, ensuring strict adherence to retrieved bilingual pairs while resolving omissions through multi-step contextual verification. This paradigm shifts from static dictionary injection to adaptive term prioritization, from rigid prompt appending to role-aware lexical integration, and from single-pass decoding to closed-loop error correction, collectively relieving data scarcity challenges in low-resource MT.

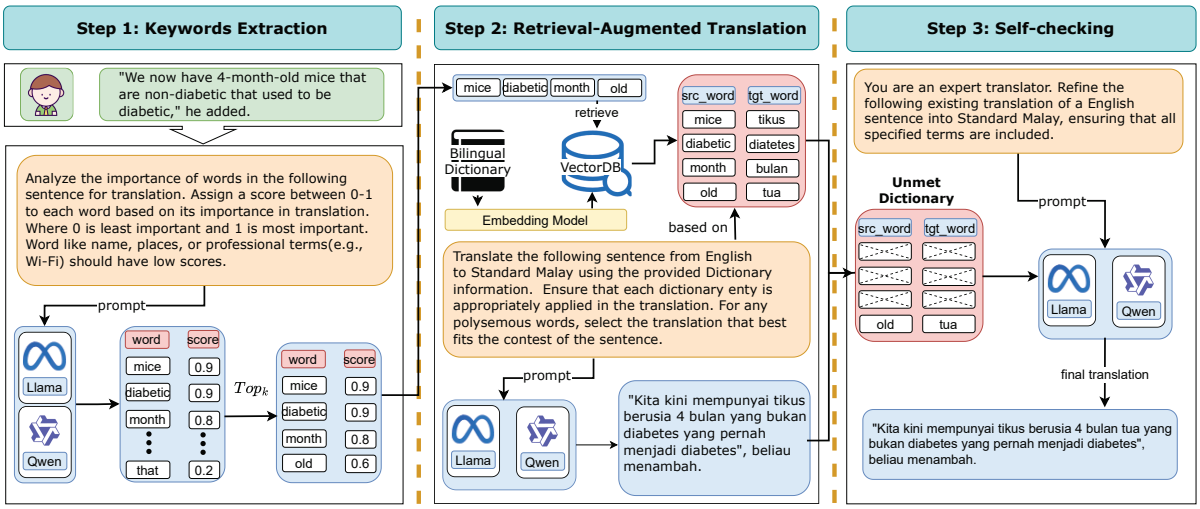


Figure 2: The proposed method of translation process.

3.2. Dictionary Extraction

In machine translation, accurately translating keywords such as domain-specific or low-frequency terms, is crucial for preserving semantic integrity. These words have a significant impact on translation quality, and errors can lead to ambiguity or loss of meaning. However, LLMs often focus on sentence sequences, neglecting the importance of these key terms. Our approach emphasizes the identification and extraction of keywords to enhance translation accuracy.

The first step in our approach involves analyzing the source sentence to identify the words that are most critical for accurate translation. We hypothesize that certain words, due to their semantic weight or specificity, play a more crucial role in preserving the intended meaning of the sentence. Given a source sentence $X = \{x_1, x_2, \ldots, x_n\}$, the LLM generates a priority score $P = \{p_1, p_2, \ldots, p_n\}$, as:

$$p_i = f_{LLM}(x_i \,|\, X,\ Prompt_k) \tag{1}$$

Keywords Evaluating Prompt

###System###
Analyze the importance of words in the following sentence for translation.
Assign a score between 0-1 to each word based on its importance in translation.
Where 0 is least important and 1 is most important.
Words like names, places, or professional terms (e.g., Wi-Fi) should have low scores.

####Human###
{sentence}

Figure 3: The prompt template for evaluating the grade of keywords.

where x_i represents individual words, our objective is to generate a priority score p_i for each word x_i in X, and $Prompt_k$ represents the prompt for evaluating the importance of keywords, as shown in Figure 3. This score reflects the importance of the word in determining the quality of the translation. f_{LLM} is the function implemented by the LLM, which takes the entire sentence X and outputs the priority score p_i for each word x_i. The LLM processes the sentence and outputs a sequence of priority scores $\{p_1, p_2, \ldots, p_n\}$, where each $p_i \in [0, 1]$ indicates the importance of word s_i. The higher the score, the more critical the word is for maintaining translation fidelity.

Ultimately, based on the above sentence $X = \{x_1, x_2, \ldots, x_n\}$ and priority score $P = \{p_1, p_2, \ldots, p_n\}$, using k as the number of top-scoring words to be selected. The goal is to select the subset $\mathcal{W}$ of k words from sentence X that have the highest scores in P. This can be mathematically described as:

$$\mathcal{W} = \{w_i \,|\, w_i \in X,\ p_i \in Top_k(P)\} \tag{2}$$

$$k = \alpha \cdot L + \sum_{i=1}^{n} \mathbb{I}(p_i \geq \beta) \tag{3}$$

Here, $Top_k(P)$ denotes the set of the top k scores from the score set P. The selected words $\mathcal{W}$ are those whose scores belong to the top k scores in the list. Our approach prioritizes influential keywords and dynamically adjusts the value of k based on keyword priority scores and input sentence length, optimizing the use of bilingual dictionary

information. The L represents the length of the input sentence, p_i is the priority score for the i-th keyword, α is a scaling factor that determines how much influence sentence length has on k, β is a threshold for the minimum priority score a keyword must meet to be included, n is the sentence length, which refers to the total number of words in a sentence. Additionally, the words are lemmatized to their base forms using the NLTK [42], which facilitates a more consistent analysis by reducing inflected forms to their root form. where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if $p_i \geq \beta$ and 0 otherwise, and n is the total number of candidate keywords.

After extracting keywords from the sentence, the next step involves retrieving their corresponding translations from a bilingual dictionary. Specifically, the dictionary $\mathcal{D}$ is defined as a set of source-target word pairs: $\mathcal{D} = \{S, T\}$, where $S = \{s_1, s_2, \cdots, s_m\}$ denotes the set of source language words and $T = \{t_1, t_2, \cdots, t_m\}$ represents the set of target language words. An embedding function $f : \mathcal{D} \rightarrow \mathbb{R}^n$ maps each word to a shared n-dimensional vector space, facilitating the computation of semantic similarities. The embeddings $\{f(d) \mid d \in \mathcal{D}\}$ are stored within a vector database to enable efficient retrieval and similarity operations.

Subsequently, for each selected keyword $w_i \in \mathcal{W}$, which resides in the source language set S, a similarity matching operation is performed against the vector database using the Maximal Marginal Relevance (MMR) [43] algorithm. The MMR algorithm balances the relevance of candidate target words to the keyword w_i with the diversity among the selected candidates, thereby mitigating redundancy. The MMR selection criterion is mathematically defined as:

$$\mathrm{MMR}(w_i, S) = \underset{s \in S}{\mathrm{argmax}}(\lambda \cdot \mathrm{Sim}(w_i, s) - (1 - \lambda) \cdot \underset{s' \in R}{\mathrm{max}}\mathrm{Sim}(\mathrm{s}, \mathrm{s}')) \tag{4}$$

$$\mathrm{Sim}(x, y) = \frac{e(x) \cdot e(y)}{\|e(x)\| \cdot \|e(y)\|} \tag{5}$$

$$\mathcal{D}_{\mathcal{W}} = \{(d_i^s,\ d_i^t) \mid d_i^s = w_i \in \mathrm{MMR}(w_i,\ S)\} \tag{6}$$

Where w_i denotes the i-th keyword selected from the subset $\mathcal{W}$ of influential words within the input sentence, λ is a scalar parameter within the range $0 < \lambda < 1$ that balances the importance of relevance versus diversity in the MMR algorithm. The similarity function $\mathrm{Sim}(x, y)$ is defined as the cosine similarity between the embedding vectors of word x and word y, thereby quantifying the semantic similarity between the two words in the vector space. Specifically, $e(x)$ and $e(y)$ are the embedding vectors of words x and y, respectively. In the MMR equation, the set R represents the subset of target words that have already been selected for inclusion in $\mathcal{D}_{\mathcal{W}}$. By considering the maximum similarity between a candidate word and the words in R, the MMR algorithm ensures that newly selected words are not only relevant to w_i but also diverse relative to the previously selected words. Ultimately, the retrieved bilingual dictionary $\mathcal{D}_{\mathcal{W}}$ is constructed by pairing each source keyword w_i with its corresponding source-target translations pair (d_i^s, d_i^t) identified through the MMR algorithm.

3.3. Retrieval-Augmented Translation

Building upon the previously established methodology for keyword selection and bilingual dictionary formation, the subsequent translation enhancement phase focuses on leveraging the tailored bilingual dictionary $\mathcal{D}_{\mathcal{W}}$ to ensure

that each dictionary entry is effectively utilized during the translation process. This is achieved by designing a prompting mechanism that integrates the entire dictionary information into the LLM-based MT pipeline, thereby enhancing translation accuracy and contextual relevance. To guarantee that every entry in the bilingual dictionary $\mathcal{D}_W$ is applied during translation, the prompt is meticulously structured to incorporate explicit instructions for the LLM.

Translation Prompt

###System###
Translate the following sentence from {source language} to {target language} using the provided Dictionary information.
Ensure that each dictionary entry is appropriately applied in the translation.
For any polysemous words, select the translation that best fits the context of the sentence.

Dictionary:{dictionary}

####Human###
{sentence}

Figure 4: The prompt template for translation based on dictionary.

The proposed translation prompt is illustrated in Figure 4, which delineates the prompt designed to guide the LLM in utilizing the bilingual dictionary $\mathcal{D}_W$ effectively. The prompt comprises three primary components: the system directive, the dictionary entries, and the source sentence. The system directive explicitly instructs the LLM to perform the translation task using the provided dictionary information. Here, {*dictionary*} is replaced with the tailored bilingual dictionary $\mathcal{D}_W$, and {*sentence*} represents the input sentence to be translated. This prompt ensures that the LLM utilizes the dictionary entries as a reference, aligning the translation output with the specified lexical mappings. To further mitigate the impact of polysemy, an additional instruction is incorporated into the system directive: *"For any polysemous words, select the translation that best fits the context of the sentence."* This modification addresses a critical limitation in dictionary-based translation, where polysemous words may have multiple possible translations. Without explicit guidance, the LLM might select a translation that does not align with the intended meaning in the given context, thereby reducing translation accuracy. By instructing the model to consider contextual appropriateness, this addition enhances lexical disambiguation and ensures that the selected translation better preserves the semantic fidelity of the source sentence.

Based on the above translation prompt, the translation process initiates by systematically integrating a bilingual dictionary, denoted as $\mathcal{D}_W$, to guide the LLM in generating an initial translation with contextual accuracy. This integration ensures that specific terminology and subtle nuances from the source language are accurately conveyed in the target language. Let Y^1 represent the initial translation output, which is obtained by conditioning the LLM with the source sentence X, the retrieved bilingual dictionary $\mathcal{D}_W$, and the translation prompt $Prompt_t$. Mathematically, this process is expressed as:

$$Y^1 = f_{LLM}(X, \mathcal{D}_W, Prompt_t) \tag{7}$$

Here, $Y^1 = \{y_1^1, y_2^1, \ldots, y_m^1\}$ denotes the sequence of tokens forming the initial translation. The function f_{LLM} represents the operation of the LLM, which processes the input data to produce the translation. The source sentence X provides

the contextual and semantic foundation for the translation, while the bilingual dictionary $\mathcal{D}_{\mathcal{W}}$ supplies relevant lexical mappings to ensure the precision and consistency of term translations. The translation prompt $Prompt_t$ includes specific instructions and contextual information that further refines the translation task.

3.4. Self-Checking

To ensure the effective utilization of all translation notes throughout the machine translation procedure, we implement an iterative prompting strategy. Although LLMs are capable of producing high-quality translations, they do not always incorporate all of the provided dictionary information in a single iteration. Consequently, we have developed an iterative mechanism in which the translation process is repeated until all constraints of the dictionary are fully satisfied.

Algorithm 1 Iterative Constrained Translation Using Large Language Models

Require: Sentence X, Pre-trained LLM, Bilingual Dictionary, Maximum Iterations N

Ensure: Final Translation Y_{final}

Extract keywords $\mathcal{W} = \{w_1, w_2, \ldots, w_k\}$ from X using LLM
Retrieve translations for keywords $\mathcal{W}$ from the Bilingual Dictionary
Initialize retrieved dictionary $\mathcal{D}_{\mathcal{W}} = \{d_1^s : d_1^t, \ldots, d_k^s : d_k^t\}$
Set current translation $T \leftarrow$ None
Initialize iteration count $n \leftarrow 0$
while $\mathcal{D}_{\mathcal{W}} \neq \emptyset$ **and** $n < N$ **do**
    Increment iteration counter: $n \leftarrow n + 1$
    Generate translation $Y_{current}$ with $\mathcal{D}_{\mathcal{W}}$ using LLM
    Check unmet constraints in $\mathcal{D}_{\mathcal{W}}$ based on $Y_{current}$
    Remove satisfied constraints from $\mathcal{D}_{\mathcal{W}}$
end while

return Final Translation Y_{final}

The Iterative Constrained Translation Using Large Language Models as shown in Algorithm 1, employs an iterative prompting strategy to enforce lexical constraints derived from a bilingual dictionary, ensuring accurate and contextually appropriate translations. The process begins by extracting key translation-relevant words from the input sentence using the LLM. These identified words are then matched with their corresponding translations retrieved from a bilingual dictionary, forming a set of lexical constraints. The LLM is prompted to generate translations while adhering to these constraints, with the process repeating iteratively to refine the output. A key challenge in iterative prompting is ensuring that all translation constraints are successfully incorporated without leading to an infinite loop. To address this, the algorithm includes a termination mechanism based on two criteria: (1) the number of iterations

must not exceed a predefined maximum N, and (2) the translation must demonstrate progress in constraint satisfaction between iterations. Specifically, after each translation attempt, the algorithm evaluates whether any remaining constraints remain unmet. If the output translation has not changed from the previous iteration or if no new constraints are satisfied, the iteration terminates early to prevent unnecessary computation. The choice of the maximum iteration count N is crucial for balancing translation accuracy and computational efficiency. A larger N allows for a more thorough refinement of the translation but increases computational costs and latency. Empirical analysis suggests that N should be set based on the complexity of the input sentence, particularly its length and the number of constrained terms. In practice, an adaptive N strategy may be employed, where the iteration limit is dynamically determined based on factors such as sentence length and the initial number of constraints. This ensures that longer sentences or those with a higher number of constraints receive sufficient iterations for refinement, while shorter sentences do not undergo unnecessary processing.

For maximum iteration count N, initially, we set $N = 10$ during early experiments and recorded the number of iterations required for each sentence in a given language pair. We found that over 70% of sentences completed translation within one iteration, and more than 90% within three iterations, indicating that most sentences reached the desired state where all dictionary information was successfully utilized. The remaining 10% of sentences either required more than three iterations or reached the maximum iteration limit without achieving perfect translation. For these cases, we retained the result from the final iteration. To balance computational efficiency and translation quality, we implement a stratified iteration strategy based on sentence complexity. Sentences with lengths less than 10 are categorized as simple sentences and assigned a maximum iteration number $N = 1$. For moderate-complexity sentences exhibiting both length less than 20 and count of keywords less than 10, we set $N = 3$. Complex sentences characterized by either length greater than 20 or count of keywords greater than 10 require more intensive processing, thus $N = 5$ is applied in these cases.

Additionally, the rationale for not employing a uniform $N = 5$ setting across all cases stems from empirical observations regarding large language models' inherent translation behavior. Preliminary experiments revealed that LLMs consistently maintain high confidence in their initial translation outputs, primarily relying on internal knowledge representations rather than external resources during the generation process. This phenomenon can be attributed to the models' extensive pre-training on multilingual corpora, which enables them to prioritize internal linguistic patterns over external lexicon resources even for simple sentences. Consequently, excessive iterations (particularly $N < 3$) on low-complexity sentences yield diminishing returns in translation accuracy while significantly increasing computational overhead. Our comparative analysis demonstrated that iterative refinements beyond $N = 1$ for length of sentence < 10 resulted in only marginal improvements (≤ 0.8 BLEU score) but consumed additional GPU memory resources. This observation underscores the critical need for adaptive N strategies that dynamically adjust iteration depth according to sentence complexity metrics. To address this challenge, our stratified approach strategically allocates computational resources through complexity-driven iteration control. By implementing progressive N thresholds based on both syntactic (sentence length) and semantic (keyword density) indicators, this adaptive mechanism effec-

tively prevents redundant iterations for structurally simple sentences while ensuring adequate refinement for complex linguistic constructs.

Self-checking Prompt

```
You are an expert translator. Refine the following existing translation of a {source language} sentence
into {target language}, ensuring that all specified terms are included..

**Original Sentence:** {original sentence}

**Existing Translation:** {current translation}

**Missing Terms:** {unmet dictionary}

**Instructions:**
- Ensure the translated sentence is fluent and natural in {target language}.
- Incorporate all the missing dictionary listed above.
- Maintain the original meaning and context of the sentence.
- For any polysemous words, select the translation that best fits the context of the sentence.
```

Figure 5: The prompt template for revised translation based on self-checking.

After generating the initial translation Y^1 , the next step involves a self-checking mechanism that refines the translation. This self-checking step uses the current translation $Y^i = \{y_1^i, y_2^i, \ldots, y_m^i\}$ and the unmet dictionary $\mathcal{D}_{\mathcal{W}_{unmet}}$ to improve the quality of the next translation Y^{i+1}. The self-checking process can be represented as:

$$Y^{i+1} = f_{LLM}(X, \mathcal{D}_{\mathcal{W}_{unmet}}, Y^i, Prompt_s) \tag{8}$$

Here, the $\mathcal{D}_{\mathcal{W}_{unmet}}$ represents the unmet dictionary containing terms that were not included in the current translation Y^i, Y^i is the current iteration of the translated sentence, and $Prompt_s$ refers to the self-checking prompt used to guide the LLM in refining the translation. These parameters work collectively to facilitate an iterative enhancement process, ensuring that each subsequent translation Y^{i+1} incorporates all specified terms while maintaining the fluency and accuracy of the translation. The self-checking prompt template is shown in Figure 5. It directs LLMs through an iterative translation refinement process by presenting the original sentence, its current translation, and a list of missing terms from the dictionary. The prompt ensures the inclusion of specified terms, maintains fluency and naturalness in the target language, preserves the original meaning and context, and selects appropriate translations for polysemous words based on contextual relevance. This approach facilitates comprehensive review and enhancement, ensuring both linguistic accuracy and contextual integrity.

The final translation is generated after the self-checking and refinement steps conclude either when the unmet dictionary $\mathcal{D}_{\mathcal{W}_{unmet}}$ becomes empty or the maximum number of iterations is reached. Thus, the final output is the refined translation obtained either by satisfying all lexical constraints through the depletion of $\mathcal{D}_{\mathcal{W}_{unmet}}$ or by reaching the predefined iteration limit.

4. Experimental Setup

Models We employed a multi-step prompting approach for keyword extraction and translation, leveraging the strengths of different models. Specifically, we utilized the *Mete-Llama-3.1-70B-Instruct* [44] and the *Qwen2-72B-Instruct* [45] for extracting keywords that have a significant impact on the quality of translation. These models are chosen for their advanced natural language understanding capabilities, which enable them to accurately extract crucial words from the source sentences. For the embedding retrieval process, we used the *bge-m3* [46] model, which is specifically designed to handle large-scale embedding tasks. This model facilitates the mapping of identified keywords to their corresponding translations in a bilingual dictionary, thereby ensuring the maintenance of semantic integrity throughout the translation process.

Table 1: The low-resource languages chosen from the FLORES-200 dataset.

Language	FLORES-200 code	Language	FLORES-200 code
Catalan	cat_Latn	**Indonesian**	ind_Latn
Croatian	hrv_Latn	**Italian**	ita_Latn
Danish	dan_Latn	**Malay**	zsm_Latn
Dutch	nld_Latn	**Norwegian**	nob_Latn
Tagalog	tgl_Latn	**Slovak**	slk_Latn

Table 2: The five selected low-resource language pairs for direct translation, bypassing English as an intermediary.

Source		Target	
Language	**FLORES-200 code**	**Language**	**FLORES-200 code**
Amharic	amh_Ethi	**Lao**	lao_Laoo
Bashkir	bak_Cyrl	**Amharic**	amh_Ethi
Buginese	bug_Latn	**Tajik**	tgk_Cyrl
Igbo	ibo_Latn	**Armenian**	hye_Armn
Kyrgyz	kir_Cyrl	**Buginese**	bug_Latn

Datasets While LLMs have shown exceptional performance in high-resource language tasks, we aim to assess their effectiveness in handling underrepresented languages, where data scarcity significantly challenges translation accuracy and quality. Our study focuses on evaluating the robustness and adaptability of LLMs in translating low-resource languages and those not encountered during training. We employed FLORES-200 devtest [15], which is a benchmark dataset for machine translation between English and low-resource languages. The creation of FLORES-200 doubles the existing language coverage of FLORES-101 [47]. FLORES-200 consists of translations from 842 distinct web articles, totaling 3001 sentences. On average, sentences are approximately 21 words long. As we aim to focus on low-resource languages, we selected 10 languages as shown in Table 1. We conducted translation experiments in both directions: from English to these languages and from these languages back to English. In addition, we selected five language pairs as shown in Table 2 for low-resource language to low-resource language translation, without using English as an intermediary. This approach further emphasizes the challenges faced when directly

translating between low-resource languages, as it bypasses the more commonly used high-resource languages like English. Meanwhile, [48] reported an issue that due to the fact that most of the corpus used for training large language models is publicly available, this indirectly leads to data pollution issues during the evaluation process of the model. To reduce this risk, we also evaluated the language pairs from the WMT22, WMT23, and WMT24 test sets, which can effectively reduce the impact of data pollution.

Evaluation Metrics For evaluation metrics, we reported the chrF++ [49] and the BLEU [50] evaluations provided by sacreBLEU[1]. We employed MTME[2] which is a simple toolkit to evaluate the performance of Machine Translation metrics on WMT test sets. Additionally, we also report COMET scores using the *Unbabel/wmt22-comet-da* model [51], specifically trained for the WMT dataset to assess translation quality based on human rankings. COMET offers a more nuanced evaluation by measuring the semantic similarity between reference and translated texts, providing a deeper understanding of how well a model captures the meaning of the input text.

Dictionary For the translation between English and low-resource languages, we utilized the ground-truth bilingual dictionaries provided by [52]. These dictionaries command [3] were meticulously crafted using Meta's internal translation tool, ensuring the accurate representation of word meanings, with particular attention given to polysemy, ensuring that multiple meanings of a single word are correctly captured. **Baseline** In the baseline experiments, we compare the performance of the current SOAT models from the FLORES-200 dataset, specifically the *NLLB-200-1.3B* and *NLLB-200-3.3B* models. These models serve as a benchmark for evaluating translation quality in a low-resource setting. Additionally, we include comparisons with similar works, such as DiPMT [8] and CoD [11], to provide a broader context for our evaluation. DiPMT employs the OPT [53] and BLOOM [3] models for dictionary-guided translation, using BLEU scores as the primary evaluation metric to assess translation quality. In contrast, CoD introduces an innovative approach by linking language dictionary information in a chain structure, leveraging ChatGPT [54] to further enhance translation accuracy and fluency. For our experiments, we employ a straightforward prompting strategy to evaluate the zero-shot translation performance in which LLMs are guided using the prompt *"Translate the following sentence from* {*source language*} *to* {*target language*}*."* This approach allows us to establish a fundamental benchmark, enabling us to compare and assess the improvements introduced by our proposed method.

Parameter Settings and Reproducibility In this section, we provided detailed information regarding the key parameters used in our method to ensure experimental reproducibility. These parameters include the scaling factor α and threshold β in Equation 3 the scalar parameter λ in Equation 4, and the maximum iteration count N in Algorithm 1. All parameters described above are shared across language pairs and are not adapted per language.

- α: The scaling factor α was determined through grid search during early experiments. We observed that the best performance improvement occurred when α was within the range of 0.3 to 0.4. To mitigate the impact of keyword omission, we selected a relatively larger value of $\alpha = 0.4$.

[1] https://github.com/mjpost/sacrebleu
[2] https://github.com/google-research/mt-metrics-eval
[3] https://github.com/facebookresearch/MUSE

- β: For β, we first conducted offline experiments where LLMs were used to evaluate priority scores for translation candidates across language pairs. These scores were saved in JSON files. Subsequently, we manually reviewed these scores and found that the majority of keywords were assigned scores above 0.6. Based on this observation, we set $\beta = 0.6$ in the final translation process.

- λ : The parameter λ, which controls the diversity of selected translation candidates in Equation 4, is adopted from the Maximal Marginal Relevance (MMR) algorithm. We followed the default setting used in previous work [43], setting $\lambda = 0.5$.

- N: The maximum iteration count N is set adaptively based on sentence complexity. To balance efficiency and quality, we follow the description in Section 3.4 and apply a stratified strategy: $N = 1$ for simple sentences (length $<$ 10), $N = 3$ for moderate sentences (length $<$ 20 and keywords $<$ 10), and $N = 5$ for complex sentences (length $\geq$ 20 or keywords $\geq$ 10).

5. Results and Discussion

In this section, we provide a detailed account of a series of experiments conducted for our proposed method, including main results, low-resource language pair translation results, contamination-free evaluation, comparative analysis of keyword identification methods, analysis and discussion of computational efficiency, ablation study and case study aimed at comprehensively evaluating the effectiveness of our proposed approach.

5.1. Main Results

This section presents the experimental results of our proposed method for enhancing low-resource MT performance in LLMs. Evaluations were conducted on the FLORES-200 benchmark for bidirectional translation between English and low-resource languages, with performance measured using BLEU and chrF++ metrics. As shown in Table 3, our method implemented with the Meta-Llama-3.1 achieves superior performance over the NLLB baseline. Specifically, it outperforms NLLB in 7 directions for English to low-resource languages translation and 8 directions for low-resource language to English translation. This improvement stems from LLM's architectural advantages which decoder-only transformer, pretrained on trillions of multilingual tokens via causal language modeling, exhibits stronger cross-lingual generalization than NLLB's encoder-decoder architecture, which relies on curated parallel data. The LLM's scale further enables finer-grained modeling of low-resource linguistic phenomena during autoregressive generation. Our method capitalizes on three key innovations to leverage these architectural strengths: 1) Active Keyword Identification: Unlike NLLB's static terminology injection, our approach dynamically detects contextually salient keywords using the LLM's emergent reasoning capabilities, improving domain-specific vocabulary handling. 2) Retrieval-Augmented Translation: We bridge lexical gaps by integrating matched bilingual dictionary pairs with the input sentence, addressing NLLB's limitations with polysemous terms and neologisms. 3) Dual-Agent

Refinement: An iterative quality control loop reduces hallucinations, mitigating the "direct translation fallacy" prevalent in single-pass encoder-decoder systems. These innovations collectively enhance semantic preservation through: Context-sensitive term prioritization, minimizing over-translation of irrelevant content; Adaptive lexical alignment, enabling dynamic updates for rare/emerging terms without retraining; Iterative error correction, improving output robustness.

We compared our method with similar approaches, including DiPMT and CoD. DiPMT reported BLEU scores using the OPT and BLOOM models, focusing on low-resource language translation tasks. In contrast, CoD used the ChatGPT model and provided both BLEU and chrF++ scores, incorporating multilingual chain-based dictionary information into the prompts to enhance translation performance. In order to reduce the gap between different models, we evaluated Qwen2 and Meta-Llama-3.1 using the method reported in their paper. The result showed that our method consistently outperforms the DiPMT method and is only weaker than the CoD method in the English to Danish direction. Compared to DiPMT and CoD, which primarily focus on maximizing the coverage of bilingual dictionary terms within the input sentence, our approach introduces a more refined strategy that prioritizes high-impact vocabulary dynamically. While ensuring comprehensive dictionary coverage can provide useful lexical constraints, it often leads to excessive prompt length, increasing computational overhead and diluting the model's focus on critical terms. In contrast, our method prioritizes keywords that exert the greatest influence on translation adequacy, allowing LLMs to allocate more attention to essential word alignments rather than processing redundant lexical constraints. Moreover, our approach integrates an iterative self-checking mechanism to further enhance translation robustness. This mechanism ensures that external knowledge provided in the prompt is effectively incorporated while mitigating the model's tendency to rely excessively on its internal priors. By continuously refining the translation output through self-verification, our method reduces hallucinations and improves terminology consistency across different contexts. This is particularly advantageous in low-resource language translation, where the absence of extensive parallel data often results in misaligned or incomplete translations.

In a comprehensive evaluation across 20 translation directions (10 English to low-resource and 10 low-resource to English translation tasks), the proposed method consistently outperformed the 0-shot baseline, yielding significant gains in translation quality. Specifically, on Qwen2, the method achieved an average improvement of +3.05 BLEU points and +3.06 chrF++ points. On Meta-Llama-3.1, the corresponding average improvements were +2.63 BLEU points and +2.01 chrF++ points. Notably, Qwen2 showed better improvement than Meta-Llama-3.1 in 13 of the 20 directions, whereas Meta-Llama-3.1 led in the remaining 7 directions. Despite its lower zero-shot baseline performance, Qwen2 benefited more from the proposed method, demonstrating larger relative gains than Meta-Llama-3.1. These results underscore that the proposed method is effective across different models and particularly beneficial for enhancing Qwen2's translation performance.

Our findings reveal that Qwen2 exhibits superior relative gains over its 0-shot baseline compared to Meta-Llama-3.1, despite its lower absolute performance. This suggests that Qwen2, while initially constrained by pretraining biases, benefits more substantially from the integration of external linguistic constraints and iterative refinement. The

Table 3: The BLEU / chrF++ scores for translation tasks across different models and methods. We report the evaluation scores of the NLLB model alongside comparable works, including DiPMT [8] and CoD [11]. For clarity, the **bold** font indicates the highest score achieved by the same model type across different methods, the notation "(−/−)" denotes the overall best system performance for a specific language pair, irrespective of the model architecture. The "Avg." column shows the average scores across all language pairs.

Model	Method	cat_Latn	hrv_Latn	dan_Latn	nld_Latn	tgl_Latn	Avg.
				English to Low-resource Languages			
NLLB-200-1.3B	0-shot	38.80 / 65.21	25.94 / 57.48	39.89 / 64.54	25.46 / 58.19	32.57 / 61.84	32.53 / 61.45
NLLB-200-3.3B	0-shot	41.38 / 66.92	(28.49 / 59.22)	41.83 / 68.97	26.36 / 58.93	(34.92 / 63.38)	34.60 / 63.28
Qwen2	0-shot	25.19 / 51.99	17.00 / 46.19	27.18 / 52.36	22.10 / 51.95	19.52 / 47.08	22.20 / 49.91
	w/DiPMT	26.43 / 53.28	18.26 / 47.63	28.14 / 53.02	23.06 / 53.23	21.82 / 48.96	23.54 / 51.22
	w/CoD	27.24 / 54.37	19.35 / 49.22	28.88 / 53.76	23.87 / 54.12	22.77 / 49.57	24.42 / 52.21
	w/ours	**30.04 / 56.56**	**21.05 / 50.54**	**30.11 / 54.60**	**25.22 / 55.59**	**24.12 / 51.63**	**26.11 / 53.78**
Meta-Llama-3.1	0-shot	38.41 / 64.58	25.31 / 55.65	42.22 / 68.68	25.17 / 57.18	30.90 / 60.73	32.40 / 61.36
	w/DiPMT	39.18 / 65.23	27.18 / 57.34	43.18 / 68.89	25.32 / 57.46	31.67 / 61.29	33.31 / 62.04
	w/CoD	41.57 / 67.02	26.93 / 57.12	(**44.74 / 70.02**)	26.03 / 58.12	33.48 / 63.12	34.55 / 63.08
	w/ours	(**42.31 / 68.31**)	**28.22 / 58.65**	44.09 / 69.71	(**27.44 / 59.30**)	**34.19 / 63.38**	(**35.25 / 63.87**)
				Low-resource Languages to English			
NLLB-200-1.3B	0-shot	46.11 / 69.71	36.84 / 64.14	47.46 / 70.56	32.51 / 60.87	46.22 / 67.82	41.83 / 66.62
NLLB-200-3.3B	0-shot	(48.04 / 70.87)	38.13 / 65.10	48.75 / 71.56	33.97 / 61.56	(48.25 / 69.27)	(43.43 / 67.67)
Qwen2	0-shot	44.59 / 68.19	35.88 / 62.25	46.26 / 69.30	31.24 / 59.49	43.21 / 65.69	40.24 / 64.98
	w/DiPMT	45.12 / 68.88	36.47 / 63.28	47.34 / 70.28	31.92 / 60.13	44.17 / 66.74	41.00 / 65.86
	w/CoD	45.78 / 69.43	36.89 / 63.92	48.02 / 71.43	32.78 / 61.04	45.38 / 67.92	41.77 / 66.75
	w/ours	**46.12 / 70.01**	**37.56 / 64.43**	**48.52 / 71.89**	**33.79 / 61.98**	**45.89 / 68.89**	**42.38 / 67.44**
Meta-Llama-3.1	0-shot	45.97 / 69.39	36.84 / 63.75	47.20 / 70.40	31.98 / 60.32	44.18 / 65.98	41.23 / 65.97
	w/DiPMT	46.43 / 69.82	37.14 / 64.85	48.88 / 72.14	32.43 / 61.85	44.62 / 66.23	41.90 / 66.98
	w/CoD	47.14 / 69.98	37.78 / 65.12	48.12 / 71.65	33.92 / 63.61	44.76 / 66.84	42.34 / 67.44
	w/ours	**47.29 / 70.23**	(**38.13 / 65.56**)	(**49.54 / 73.14**)	(**34.28 / 63.93**)	**45.28 / 67.60**	**42.90 / 68.09**

Model	Method	ind_Latn	ita_Latn	msa_Latn	nob_Latn	slk_Latn	Avg.
				English to Low-resource Languages			
NLLB-200-1.3B	0-shot	44.56 / 70.58	28.62 / 61.41	39.24 / 68.30	30.47 / 60.99	29.63 / 59.34	34.50 / 64.12
NLLB-200-3.3B	0-shot	45.59 / 71.3	29.94 / 62.07	(40.94 / 69.54)	31.99 / 62.23	31.97 / 61.37	36.09 / 65.30
Qwen2	0-shot	39.42 / 65.41	27.56 / 57.26	30.43 / 59.95	19.32 / 46.33	15.68 / 42.36	26.48 / 54.26
	w/DiPMT	41.23 / 67.86	28.98 / 58.87	31.23 / 60.46	21.89 / 48.52	17.43 / 44.86	28.15 / 56.11
	w/CoD	42.75 / 68.32	30.62 / 60.03	32.77 / 62.14	22.34 / 49.08	17.89 / 45.23	29.27 / 56.96
	w/ours	**43.88 / 69.29**	**31.73 / 60.72**	**33.81 / 63.36**	**23.26 / 50.56**	**19.20 / 46.26**	**30.38 / 58.04**
Meta-Llama-3.1	0-shot	42.62 / 69.50	28.52 / 60.48	36.78 / 65.82	29.58 / 59.87	27.47 / 58.13	32.99 / 62.76
	w/DiPMT	44.15 / 71.38	29.03 / 61.07	38.24 / 67.35	30.88 / 61.04	29.32 / 60.84	34.32 / 64.34
	w/CoD	43.29 / 70.88	30.83 / 62.15	37.82 / 66.43	31.43 / 61.79	29.87 / 61.15	34.65 / 64.48
	w/ours	(**46.85 / 73.23**)	(**31.19 / 62.57**)	**38.88 / 67.98**	(**32.70 / 62.64**)	(**32.06 / 62.36**)	(**36.34 / 65.76**)
				Low-resource Languages to English			
NLLB-200-1.3B	0-shot	42.98 / 68.66	34.02 / 61.28	44.03 / 69.0	39.99 / 65.28	38.69 / 65.42	39.94 / 65.93
NLLB-200-3.3B	0-shot	45.21 / 70.12	34.68 / 61.95	46.05 / 70.07	42.72 / 67.05	40.32 / 66.43	41.80 / 67.12
Qwen2	0-shot	43.83 / 68.51	32.98 / 60.17	43.43 / 67.06	41.84 / 66.43	37.70 / 63.89	39.96 / 65.21
	w/DiPMT	44.23 / 69.15	33.56 / 61.48	45.78 / 68.43	42.34 / 67.85	38.42 / 64.53	40.87 / 66.29
	w/CoD	45.06 / 69.73	34.63 / 62.56	45.92 / 69.01	42.78 / 68.23	38.88 / 65.04	41.45 / 66.91
	w/ours	**45.36 / 70.02**	**34.71 / 62.82**	**46.11 / 69.52**	**43.25 / 68.77**	**39.02 / 65.66**	**41.69 / 67.36**
Meta-Llama-3.1	0-shot	44.18 / 68.76	33.81 / 61.16	43.40 / 67.85	42.99 / 67.51	38.11 / 64.40	40.50 / 65.94
	w/DiPMT	46.38 / 70.23	34.46 / 62.15	45.34 / 69.26	43.86 / 68.97	38.86 / 65.34	41.78 / 67.19
	w/CoD	45.68 / 69.54	35.02 / 62.98	44.48 / 68.88	44.52 / 69.74	39.43 / 66.52	41.83 / 67.53
	w/ours	(**47.23 / 71.33**)	(**35.28 / 63.44**)	(**46.91 / 70.94**)	(**45.88 / 70.62**)	(**40.47 / 67.07**)	(**43.15 / 68.68**)

model's ability to internalize and adapt to lexical guidance enables more effective mitigation of pretraining deficiencies, resulting in a steeper performance gain relative to its zero-shot translation capabilities. In contrast, Meta-Llama-3.1, with its stronger baseline performance, exhibits a smaller performance improvement, indicating that models with higher initial proficiency may derive relatively smaller gains from post-hoc constraint-based refinements. Further analysis reveals that Qwen2's heavy reliance on Chinese corpora during pretraining introduces inherent biases that manifest as cross-lingual interference in zero-shot scenarios, particularly for typologically distant low-resource languages. This data imbalance initially leads to suboptimal lexical choices and unintended code-switching tendencies in non-Chinese target languages and we will further analyze this situation in Section 5.7. However, the proposed method's integration of explicit linguistic constraints effectively mitigates these biases by enforcing target-language structural alignment and suppressing hallucinated content. Crucially, Qwen2 demonstrates enhanced capacity to internalize external linguistic priors during constrained decoding cycles, enabling progressive error correction even when initial translations diverge significantly from target norms. This adaptability proves particularly impactful in morphologically complex languages, where the model compensates for pretraining deficiencies by dynamically incorporating domain-specific rules through multi-stage verification.

5.2. Low-Resource Language Pair Translation Results

Table 4 presents the chrF++ scores for low-resource to low-resource language translation tasks using the Meta-Llama-3.1 and Qwen2 models. Three translation strategies were assessed: **0-shot** where the models perform direct translation without external guidance; **D** where **D**ictionary information is incorporated into the translation process; and **D+S** which combines **D**ictionary information with the **S**elf-checking mechanism to refine translations.

Table 4: The chrF++ scores for low-resource to low-resource language translation tasks using Meta-Llama-3.1 and Qwen2 models. Three translation strategies are evaluated: 0-shot (direct translation without additional information), **D** (incorporating **D**ictionary information), and **D+S** (combining **D**ictionary information with the **S**elf-checking mechanism). The results highlight the effectiveness of combining dictionary constraints and self-checking in improving translation quality for low-resource language pairs.

Language Pair	**Meta-Llama-3.1**			**Qwen2**		
	0-shot	**w/D**	**w/D+S**	**0-shot**	**w/D**	**w/D+S**
amh_Ethi → lao_Laoo	13.86	14.58	**16.74**	11.52	10.26	**12.58**
bak_Cyrl → amh_Ethi	10.11	13.47	**14.62**	2.99	4.62	**7.78**
bug_Latn → tgk_Cyrl	23.14	23.78	**24.62**	6.88	7.62	**8.31**
ibo_Latn → hye_Armn	16.81	18.24	**19.47**	8.89	**10.12**	9.78
kir_Cyrl → bug_Latn	23.71	24.53	**24.68**	12.61	13.23	**14.18**
System Average:	17.53	18.91	**20.03**	8.58	9.16	**10.53**
System Wins:	0/5	0/5	5/5	0	1/5	4/5

The results demonstrate that incorporating dictionary information **(D)** significantly improves the translation quality for both models compared to the 0-shot baseline. Specifically, the Meta-Llama-3.1 model achieves an average chrF++ score improvement from 17.53 **(0-shot)** to 18.91 **(D)** and further to 20.03 when using the **(D+S)** strategy. This

highlights the combined benefits of lexical constraints and self-checking in enhancing translation accuracy. Moreover, Meta-Llama-3.1 achieves the highest chrF++ scores across all five language pairs under the **(D+S)** framework. For the Qwen2 model, the **D+S** strategy also yields substantial performance improvements, with the average chrF++ score increasing from 8.58 **(0-shot)** to 10.53. Notably, Qwen2 achieves the highest scores in four out of five language pairs when using **(D+S)**, showcasing its potential in low-resource translation tasks with additional guidance.

These findings underscore the importance of integrating external linguistic knowledge, such as dictionary information, and leveraging self-checking mechanisms to address the challenges of low-resource language translation. While both models benefit from these strategies, Meta-Llama-3.1 demonstrates superior overall performance, particularly under the **(D+S)** framework.

5.3. Contamination-Free Evaluation

Moreover, experiments were conducted using the most recent machine translation evaluation datasets, specifically WMT22, WMT23, WMT24. These datasets provide a valuable benchmark for evaluating the performance of LLMs in translation tasks. One of the key advantages of utilizing these updated datasets is their ability to mitigate the issue of data contamination that can arise during the training process. Data contamination refers to the inadvertent encounter of portions of the test set by models during training, which can result in performance scores that are artificially inflated and do not accurately reflect the model's true capabilities in a real-world setting.

Table 5: The BLEU / COMET score with WMT testset using Meta-Llama3.1 and Qwen2.

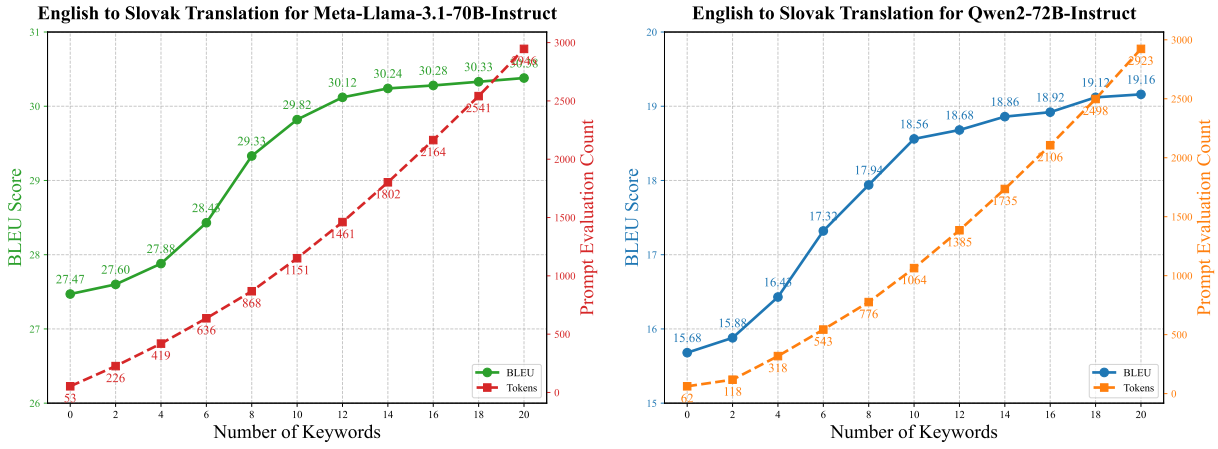
Dataset	Language Pair	**Meta-Llama-3.1**		**Qwen2**	
		Baseline	Ours	Baseline	Ours
WMT22	en-de	32.51 / 86.15	**33.92 / 86.45**	26.34 / 84.12	**28.45 / 85.14**
	en-hr	23.88 / 86.32	**26.09 / 87.75**	14.96 / 79.02	**18.42 / 84.63**
WMT23	en-de	38.69 / 82.51	**40.13 / 82.72**	34.97 / 80.45	**36.28 / 81.56**
	en-he	22.54 / 81.60	**24.26 / 82.74**	7.90 / 66.85	**16.78 / 78.53**
WMT24	en-de	29.42 / 80.21	**30.79 / 80.86**	24.58 / 77.48	**25.64 / 79.02**
	en-es	41.81 / 81.64	**42.64 / 81.81**	37.55 / 79.84	**38.71 / 81.25**

By incorporating WMT22-24, we ensured that the evaluation process was more robust and free from the potential issues associated with overfitting to previously seen data. As shown in Table 5, the results from our experiments using these newer datasets demonstrated a consistent improvement in translation quality across various language pairs. The experimental results demonstrate consistent performance improvements across multiple configurations. For the Meta-Llama-3.1 model, the proposed method achieves absolute BLEU score gains ranging from +0.90 to +2.21 across high-resource language pairs (en-de, en-es), with the most significant improvement observed in WMT22 en-de (+1.41 BLEU). COMET scores show stable enhancements from 0.30 to 1.43 points, indicating improved translation quality coherence. Notably, low-resource pairs exhibit more pronounced gains: en-hr (WMT22) shows +2.21 BLEU

improvement, while en-he (WMT23) achieves +1.72 COMET increase, suggesting the method's effectiveness in data-scarce scenarios. The Qwen2 model displays even stronger relative improvements. The Baseline-to-Ours comparisons reveal BLEU score increments of 2.11–8.88 points, with particularly remarkable progress in challenging low-resource settings: en-he (WMT23) achieves the 112.4% relative improvement (+8.88 BLEU), accompanied by a substantial +11.68 COMET gain. The exceptional improvements observed with the Qwen2 model can be attributed to its initial struggle with zero-shot translation, where it frequently generated substantial non-target language hallucinations and our method significantly mitigated these issues. High-resource pairs maintain competitive enhancements, exemplified by en-es (WMT24) with +1.16 BLEU and +1.41 COMET improvements. Cross-model analysis shows Meta-Llama-3.1 generally outperforms Qwen2 in absolute scores (e.g., WMT24 en-de: 30.79 for Meta-Llama-3.1 vs. 25.64 for Qwen2), though Qwen2 exhibits greater relative gains in low-resource conditions. The upward trend in performance across all experimental conditions underscores the effectiveness of these datasets in providing a cleaner and more reliable evaluation framework. This improvement serves to validate the robustness of our methodology and also highlights the importance of using up-to-date and diverse data for evaluating LLMs in machine translation tasks.

These results underscore the importance of addressing data contamination in training large-scale models, as reliance on outdated or overly familiar datasets can obscure the true performance capabilities of the models. By employing the latest evaluation datasets, we were able to obtain a more accurate and meaningful assessment of translation quality, thereby providing a clearer picture of the improvements brought about by our approach.

5.4. Comparative Analysis of Keyword Identification Methods



(a) Meta-Llama-3.1-70B-Instruct

(b) Qwen2-72B-Intruct

Figure 6: Performance analysis of keyword selection strategies with English to Slovak Translation task.

In order to evaluate the impact of different keyword identification methods, four approaches were compared:

- **0-shot**: Translations without word constraints, which corresponds to a zero-shot translation approach.

- **fix-10**: Fixing the number of words extracted from the source sentence, referred to as "fix-10" in the table, where we select exactly 10 words from the source sentence.
- **random-10**: Randomly selecting words from the source sentence, denoted as "random-10" in the table, where we use a random function to select 10 words from the source sentence.
- **LLM-Based**: Using the LLM to identify important words in the source sentence, referred to as "LLM-based" in the table, which is the method proposed in this paper.

Table 6: BLEU scores for different translation methods by keywords identification across various language pairs, the **bold** font indicates the method of obtaining the highest score.

Model	Method	cat_Latn	hrv_Latn	dan_Latn	nld_Latn	tgl_Latn	ind_Latn	ita_Latn	msa_Latn	nob_Latn	slk_Latn
English to low-resource languages											
Qwen2	0-shot	25.19	17.00	27.18	22.10	19.52	39.42	27.56	30.43	19.32	15.68
	fix-10	27.35	18.52	28.93	23.57	20.14	40.52	29.00	31.24	20.52	16.81
	random-10	28.76	20.37	29.75	24.86	21.63	42.73	30.58	32.92	22.33	18.56
	LLM-Based	**30.04**	**21.05**	**30.11**	**25.22**	**24.12**	**43.88**	**31.72**	**33.81**	**23.26**	**19.20**
Meta-Llama-3.1	0-shot	38.41	25.31	42.22	25.17	30.90	44.62	28.52	36.78	29.58	27.47
	fix-10	39.85	26.12	43.54	26.38	31.25	45.23	29.40	37.53	30.65	28.32
	random-10	41.21	27.86	43.95	26.92	32.75	45.86	30.78	38.22	31.54	29.82
	LLM-Based	**42.31**	**28.22**	**44.09**	**27.44**	**34.19**	**46.85**	**31.19**	**38.88**	**32.70**	**32.06**
Low-resource languages to english											
Qwen2	0-shot	44.59	35.88	46.26	31.24	43.21	43.83	32.98	43.43	41.84	37.70
	fix-10	45.11	36.72	46.78	32.69	44.56	44.95	33.51	44.23	42.55	38.23
	random-10	45.78	36.64	47.48	33.34	45.74	44.62	34.28	45.73	43.08	38.74
	LLM-Based	**46.12**	**37.56**	**48.52**	**33.79**	**45.89**	**45.36**	**34.71**	**46.11**	**43.25**	**39.02**
Meta-Llama-3.1	0-shot	45.97	36.84	47.20	31.98	44.18	44.18	33.81	43.40	42.99	38.11
	fix-10	46.63	37.65	48.37	33.14	44.73	45.06	34.54	44.57	44.43	39.29
	random-10	47.13	37.96	49.21	34.03	45.08	46.75	34.89	46.51	45.45	40.28
	LLM-Based	**47.29**	**38.13**	**49.54**	**34.28**	**45.28**	**47.23**	**35.28**	**46.91**	**45.88**	**40.47**

To validate the effectiveness of LLM-based keyword identification, we focus on introducing the following two methods: (1) fix-10, a fixed strategy for extracting the first 10 words from a source sentence. This method serves as a control to assess the impact of context-agnostic keyword selection. And (2) random-10, a random strategy where 10 words are selected uniformly at random from the source sentence. This evaluates the necessity of semantic-aware keyword prioritization. The choice of "10 keywords" was determined through preliminary experiments, where we observed that increasing keyword counts beyond 10 led to diminishing returns in translation quality while significantly increasing computational overhead. We conducted preliminary experiments on the English → Slovak translation task, varying the number of keywords from 5 to 20. Figure 6(a) and Figure 6(b) show that BLEU scores plateau at 10 keywords, while computational cost (tokens that prompt evaluation count) increases linearly.Selecting 10 keywords

achieves a Pareto-optimal balance between translation quality (retaining 97.6% of peak BLEU performance) and efficiency (reducing 42% of computational overhead compared to 20 keywords). By comparing the design with fix and random methods, we aim to further demonstrate the performance improvement brought about by our approach.

As shown in Table 6, our LLM-based approach significantly outperforms both the fixed and random extraction methods, as well as the unconstrained translation approach. This demonstrates the efficacy of using a large language model to identify contextually important words in the source sentence, providing more targeted and relevant word constraints during the translation process. The fixed extraction method, while offering a controlled selection of words, does not account for the varying importance of different words in the sentence, potentially leading to suboptimal translations. Similarly, the random extraction method introduces unnecessary variability, as the selected words may not be those that contribute most significantly to translation quality. In contrast, the LLM-based approach dynamically adapts to the linguistic context, identifying key terms that are crucial for accurate translation. As a result, it yields higher-quality translations.

5.5. Analysis and Discussion of Computational Efficiency

This study quantitatively evaluates the impact of our proposed method (both with and without the iterative self-checking mechanism) by analyzing its performance in terms of computational efficiency and translation quality compared to existing baseline methods. This detailed analysis seeks to substantiate our claims and clearly elucidate the contributions of each component of our approach. Due to the addition of an iterative self-checking mechanism, our method does indeed have multiple calls to LLMs, resulting in increased computational overhead, as we will no longer count their input tokens. To comprehensively understand computational efficiency, this study introduces two key metrics:

- **Input Tokens (tokens/it)**: This metric quantifies the average number of tokens processed per translation item by Large Language Models (LLMs) during inference. It includes both prompt tokens and source text tokens. Lower input token values indicate more efficient prompt engineering, as the model requires less contextual information to generate translations, thereby reducing the computational load associated with token processing.
- **Time Cost (s/it)**: This metric measures the average wall-clock time required to generate one translated output, representing the end-to-end latency from input submission to completed translation. It encompasses computation time, memory access overhead, and decoding time. Time cost is a critical indicator for real-time applications, where low latency is essential for user experience and system responsiveness.

The data presented in Table 7 provides strong evidence for the computational efficiency of our proposed dictionary-driven single-pass translation method ("Ours w/D") compared to DiPMT and CoD. Firstly, in terms of input token efficiency, "Ours w/D" consistently processes fewer input tokens per translation item across both Qwen2-72B-Instruct and Meta-Llama-3.1-70B-Instruct models. For instance, in the English-to-non-English direction for Qwen2, "Ours

Table 7: Computational Efficiency and Translation Quality Comparison across Methods. "Ours w/D" denotes our method using dictionary information without self-checking, forming a single-pass translation comparable to DiPMT and CoD. "Ours w/D+S" incorporates dictionary information with iterative self-checking. BLEU scores, average input tokens per item (tokens/it), and average time cost per item (s/it) are reported for English to low-resource languages (en→xx) and low-resource languages to English (xx→en) directions.

Method	**en→xx**			**xx→en**		
	BLEU	input tokens (tokens/it)	time cost (s/it)	BLEU	input tokens (tokens/it)	time cost (s/it)
Qwen2-72B-Instruct						
DiPMT	25.85	348	12.27	40.94	336	13.12
CoD	26.85	693	14.66	41.61	623	15.87
Ours w/D	26.99	319	11.82	41.71	298	12.65
Ours w/D+s	28.24	-	18.84	42.03	-	19.28
Meta-Llama-3.1-70B-Instruct						
DiPMT	33.82	283	5.61	41.84	273	6.38
CoD	34.60	684	6.80	42.09	692	8.64
Ours w/D	34.84	278	5.42	41.92	285	6.16
Ours w/D+s	35.79	-	8.52	43.02	-	9.12

w/D" uses only 319 tokens/it, significantly less than DiPMT's 348 tokens/it, and markedly lower than CoD's 693 tokens/it. Similar trends are observed across all language pairs and models. This empirically confirms our assertion that our method, through a refined strategy of dynamically prioritizing high-impact vocabulary, avoids the "overly lengthy prompts" issue that can "increase computational overhead while diluting focus on critical terms". Our LLM-based keyword identification method effectively filters for the most relevant lexical constraints, leading to more concise and efficient prompts. In addition, regarding reduced time cost, "Ours w/D" demonstrates lower average wall-clock time. For the English to low-resource languages direction with Qwen2, "Ours w/D" takes 11.82 s/it, outperforming DiPMT (12.27 s/it) and CoD (14.66 s/it). Meta-Llama-3.1 also shows "Ours w/D" as the most efficient single-pass translation method (5.42 s/it for English to low-resource languages). The direct correlation between reduced input tokens and lower time cost highlights the practical efficiency gains of our method, making it more suitable for scenarios requiring rapid inference. Finally, in terms of translation quality, despite the reduced computational overhead, "Ours w/D" maintains or slightly improves BLEU scores compared to DiPMT and CoD. For example, on Qwen2's English to low-resource languages, "Ours w/D" achieves 26.99 BLEU, surpassing DiPMT (25.85) and CoD (26.85). This indicates that our method effectively focuses the LLM's attention on critical terms without "diluting focus," thereby achieving competitive or superior translation quality with higher efficiency. The observed efficiency gains of "Ours w/D" over DiPMT and CoD are not merely incremental improvements but highlight a fundamental difference in prompt engineering philosophy. Our method's "refined strategy that prioritizes high-impact vocabulary dynamically" suggests that carefully curated, high-quality context is more effective than simply voluminous context. This finding has significant implications for broader principles of interacting with LLMs: for specific tasks like machine translation, intelligent filtering and prioritization of input information can lead to superior performance and efficiency, challenging the intuitive "more data is better" prompting approach.

As anticipated, the "Ours w/D+S" method incurs a significantly higher time cost compared to "Ours w/D". For instance, the time cost for Qwen2 English-to-non-English jumps from 11.82 seconds/item for "Ours w/D" to 18.84 seconds/item for "Ours w/D+S", an increase of approximately 60%. A similar increase is observed for Meta-Llama-3.1 (e.g., from 5.42 s/it to 8.52 s/it for English to low-resource language). This increased latency is a direct consequence of the iterative nature of the self-checking process , where the model performs multiple inference calls to refine its output. Our method explicitly acknowledges this trade-off, stating that "a larger N allows for a more thorough refinement of the translation but increases computational costs and latency". Regarding the input tokens for the iterative process, the "input tokens" metric for "Ours w/D+S" is marked with "-", indicating that a single static value is not directly applicable or easily comparable to single-pass methods. The iterative self-checking mechanism involves multiple rounds of prompting, where the model processes the original source sentence, the current translation, and unmet dictionary constraints. Therefore, the total input tokens for "Ours w/D+S" would be the accumulation of these iterations, making "time cost" a more representative end-to-end metric for actual latency and overall computational load. Despite the increased computation time, "Ours w/D+S" consistently yields higher BLEU scores. This demonstrates that the iterative refinement process brings substantial quality benefits. Ablation studies further confirm this, showing that the self-checking mechanism "significantly enhances translation quality" by "ensuring consistency, reducing errors, and increasing overall accuracy" through "verification and adjustment of its outputs". The significantly increased time cost of "Ours w/D+S" might initially be perceived as a drawback. However, the consistent and notable improvement in BLEU scores (and the observed ability to mitigate hallucinations in case studies) justifies this overhead. This highlights a critical design consideration in LLM-driven applications: for tasks where accuracy, consistency, and robustness are paramount (e.g., professional translation, domain-specific contexts, or low-resource languages with poor initial quality), investing additional computational cycles for iterative refinement is a worthwhile trade-off. This positions the "Ours w/D+S" method as a high-quality, robust solution for demanding scenarios, even if it is not the fastest.

5.6. Ablation Study

In addition, we also explored the effects of different translation strategies, including the 0-shot translation approach, using only a translation dictionary without the Self-Checking mechanism, and incorporating the Self-Checking mechanism into the translation process. The results as shown in Figure 7, indicate that while the use of a translation dictionary without the Self-Checking mechanism leads to an improvement in BLEU scores compared to the baseline (0-shot), it still falls short of the performance achieved when the Self-Checking mechanism is applied.

The observed improvement with the translation dictionary alone can be attributed to the additional lexical constraints provided by the dictionary, which helps guide the translation model towards more accurate word choices. However, this approach does not fully capture the context and intricate nuances of sentence-level meaning. Without the Self-Checking mechanism, the translation process remains somewhat static, as it does not involve any form of post-translation evaluation or correction. The model may still produce translations that are syntactically or semantically

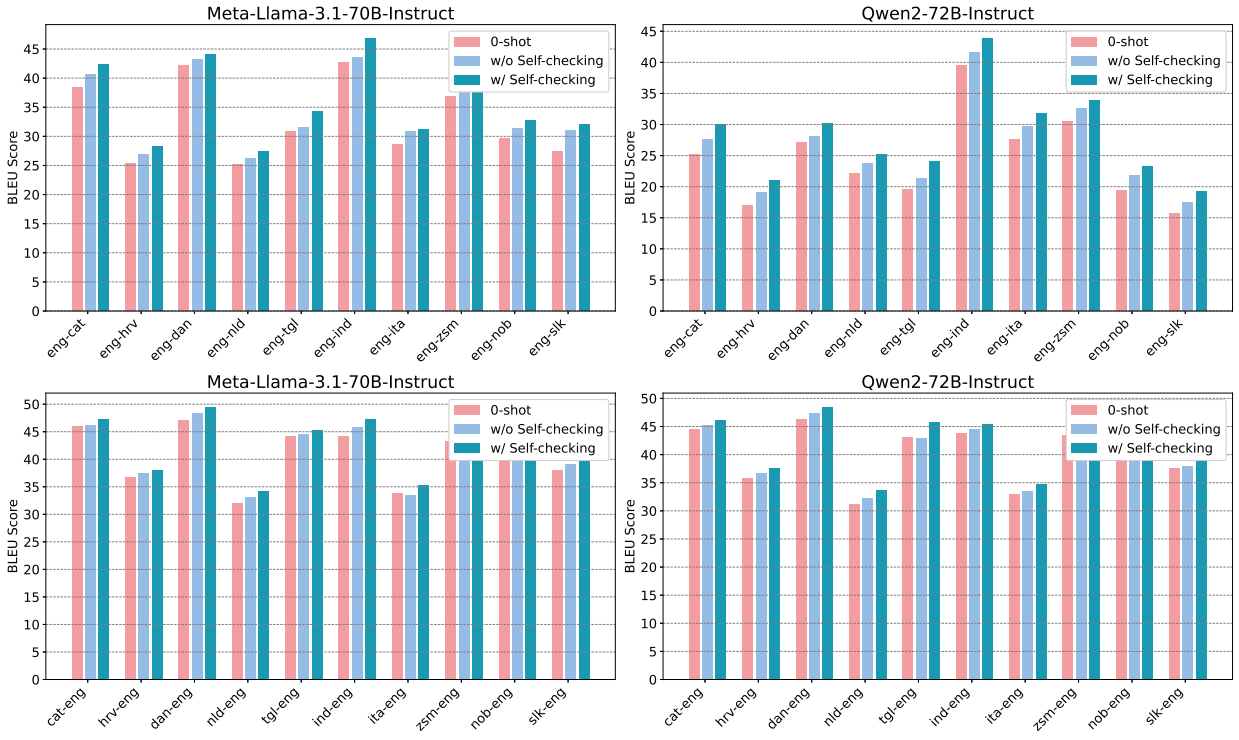


Figure 7: Comparison of BLEU scores for Meta-Llama-3.1-70B-Instruct and Qwen2-72B-Instruct with and without self-checking versus 0-shot.

flawed, especially when dealing with ambiguous or polysemous terms, which are common in low-resource language pairs. On the other hand, the inclusion of the Self-Checking mechanism significantly enhances the translation quality. This mechanism allows the model to verify and adjust its outputs after the initial translation, ensuring consistency, reducing errors, and increasing overall accuracy. The ability of the Self-Checking mechanism to dynamically refine the translation output based on context makes it particularly effective in addressing issues such as mistranslations or poor lexical choices, which are more prevalent when only a dictionary is used. Furthermore, Self-Checking helps to maintain fluency and grammatical correctness, which contributes to the higher BLEU scores observed with this approach.

In summary, our ablation study highlights the superiority of the LLM-based keyword identification method combined with the Self-Checking mechanism. The LLM-based approach significantly outperforms both fixed and random word extraction methods, as well as the zero-shot translation approach, by dynamically identifying contextually important words for more accurate translations. Furthermore, the integration of the Self-Checking mechanism further enhances translation quality by allowing the model to refine its output and correct errors, ensuring greater consistency and fluency. Together, these innovations demonstrate the effectiveness of our method in improving machine translation performance.

5.7. Case Study

The case study as shown in Figure 8 examines the translation quality of Meta-Llama-3.1 for English-to-Slovak tasks, focusing on the impact of incorporating dictionary constraints and self-checking mechanisms. The source sentence and its corresponding reference translation highlight the challenges of accurately translating low-resource sentences, particularly when specific keywords significantly influence the overall translation quality.

Meta-Llama-3.1 (translation from English to Slovak)	
Source Sentence	No extra transport is being put on and overground trains will not stop at Wembley, and car parking and park-and-ride facilities are unavailable at the ground.
Refence	K dispozícii nie je žiaden ďalší druh dopravy, nadzemné vlaky nezastavujú vo Wembley a taktiež nie sú k dispozícii parkoviská ani zariadenia s možnosťou park-and-ride.
Keywords	["transport", "trains", "parking", "facilities", "ground"]
Dictionary	{'transport': ['dopravy', 'dopravné', 'transport'], 'trains': ['vlaky', 'vlakov'], 'parking': ['parkoviská, 'parkovanie'], 'facilities': ['vybavenie', 'zariadenia'], 'ground': ['mleté', 'pozemné', 'zemi', 'zem', 'pozemný']}
Zero-Shot Output BLEU=3.39 chrF++= 27.74	Žiadne ďalšie dopravné služby nebudú zavedené a nadzemné vlaky sa neučia zastavení na stanici Wembley, a parkovacie plochy a služba park and ride budú uzamknuté pri štadióne.
w/ Dictionary BLEU=8.69 chrF++= 46.38	Nie sú k dispozícii žiadne ďalšie prepravy a nadzemné vlaky sa nezastavujú na Wembley, a parkoviská áut a park-and-ride zariadenia nie sú dostupné na štadióne.
Unmet Dictionary	{'transport': ['dopravy', 'dopravné'], 'ground': ['mleté', 'pozemné', 'zemi', 'zem', 'pozemný']}
w/ Dictionary + Self-Checking BLEU: 8.86 chrF++:47.38	Nie sú k dispozícii žiadne ďalšie dopravy a nadzemné vlaky sa nezastavujú na Wembley, a parkoviská áut a park-and-ride zariadenia nie sú dostupné na pozemný štadión.

Figure 8: A case study with Meta-Llama-3.1 model for translation task from English to Slovak.

In the zero-shot translation scenario, where the model directly generates translations without external guidance, the BLEU score is 3.39, and the chrF++ score is 27.74. The output demonstrates limited accuracy, failing to capture the semantic and lexical nuances of the reference sentence. Keywords such as “transport”, “trains”, and “ground” are inadequately translated, leading to a significant loss in meaning.By integrating dictionary constraints, the BLEU score improves to 8.69, and the chrF++ score increases to 46.38. This improvement underscores the importance of providing the model with explicit lexical information. Keywords such as “vlaky” (trains) and “parkovanie” (parking) are accurately translated, resulting in better alignment with the reference sentence. However, certain contextual nuances remain unresolved, as indicated by discrepancies in the translation of “ground” and “facilities”. To address these issues, the addition of a self-checking mechanism further enhances translation performance. With the combined use of dictionary constraints and self-checking, the BLEU score reaches 8.86, and the chrF++ score improves slightly

to 47.38. This approach ensures a higher degree of consistency and semantic accuracy, particularly for critical terms such as “dopravný” (transport) and “pozemný” (ground). The self-checking mechanism effectively resolves lexical ambiguities and ensures the integrity of the final translation.

Qwen2 (translation from English to Norwegian)	
Source Sentence	Duvall, who is married with two adult children, did not leave a big impression on Miller, to whom the story was related.
Refence	Duvall, som er gift og har 2 voksne barn, hadde liten innvirkning på Miller, som historien var relatert til.
Keywords	[“married”, "adult", "children", "impression", "story", "related"]
Dictionary	{'married': ['gift'], 'adult': ['voksen', 'voksent', 'voksne'], 'children': ['barn'], 'impression': ['inntrykk'], 'story': ['hendelse', 'story', 'historien'], 'related': ['relatert', 'relaterte']}
Zero-Shot Output BLEU=26.19 chrF++= 49.65	Duvall, som er gift med to voksne barn,留下来et ikke et stort inntrykk på Miller, til hvem historien ble relatert.
w/ Dictionary BLEU=26.97 chrF++= 55.07	Duvall, som er gift med to vokne barn, gav ikke et stort inntrykk på Miller, til hvem historien var relatert.
Unmet Dictionary	{ }
w/ Dictionary + Self-Checking BLEU=26.97 chrF++= 55.07	Duvall, som er gift med to vokne barn, gav ikke et stort inntrykk på Miller, til hvem historien var relatert.

Figure 9: A case study with Qwen2 model for translation task from English to Norwegian.

For the case study of Qwen2 model is shown in Figure 9, the model’s performance for English-to-Norwegian translation is analyzed, highlighting challenges in zero-shot translation. In the zero-shot scenario, the model unexpectedly introduced non-target language content, outputting Chinese characters instead of Norwegian, which indicates a hallucination issue. The BLEU and chrF++ scores for zero-shot translation were 26.19 and 49.65, respectively. By incorporating dictionary constraints, the model achieved improved accuracy, with BLEU increasing to 26.97 and chrF++ to 55.07, correctly translating critical terms such as “gift” (married) and “voksne barn” (adult children). Adding a self-checking mechanism maintained these improvements, ensuring consistent translations aligned with the reference. This study highlights the limitations of zero-shot translation in Qwen2 and demonstrates the effectiveness of dictionary constraints and self-checking in mitigating hallucination issues and enhancing translation quality.

Meanwhile, we also added Table 8 that compares with DiPMT and CoD by Qwen2-72B-Instruct. As shown in Table 8, DiPMT and CoD overload prompts with exhaustive term mappings, causing the LLM to prioritize low-relevance

Table 8: A case study that comparison with DiPMT and CoD by Qwen2-72B-Instruct.

Source Sentence (English)	Travelling by plane can be a scary experience for people of all ages and backgrounds, particularly if they've not flown before or have experienced a traumatic event.
Target Sentence (Slovak)	Cestovanie lietadlom môže byť strašidelným zážitkom, bez ohľadu na vek či povôd cestujúcich. A to najmä v prípade, ak nikdy predtým neleteli alebo zažili traumatizujúcu udalosť.
DiPMT's Prompt	"traumatic" means "traumatické". "if" means "ak". "people" means "Ľudia". "event" means "udalosť". "all" means "všetky". "flown" means "vyletel". "or" means "alebo". "backgrounds" means "pozadie,". "and" means "a)". "for" means "pre". "particularly" means "najmä". "a" means "a)". "experience" means "skúsenosti". "scary" means "strašidelný". "by" means "v". "plane" means "lietadlo". "have" means "majú". "before" means "pred". "not" means "nie". "Travelling" means "Cestovanie". "experienced" means "skúsený". "of" means "z". "be" means "byť". "can" means "môže". "they've" means "Oni majú". "ages" means "vek". Translate the following text from English to Slovak: {Source Sentence}
CoD's Prompt	"traumatic" means "traumatické" means "traumatisant" means "traumático" means "Traumatische". "if" means "ak" means "si le" means "si el" means "wenn". "people" means "Ľudia" means "Les personnes" means "personas" means "Menschen". "event" means "udalosť" means "événement" means "El evento" means "Veranstaltung". "all" means "všetky" means "tout" means "todos" means "alle". "flown" means "vyletel" means "volé" means "No se puede" means "Flüge". "or" means "alebo" means "ou" means "o el" means "oder". "backgrounds," means "pozadie," means "les antécédents," means "los antecedentes," means "Hintergründe". "and" means "a)" means "et de" means "y el" means "und". "for" means "pre" means "pour" means "para el" means "für". "particularly" means "najmä" means "particulièrement" means "En particular," means "Besonders". "a" means "a)" means "le" means "a) El" means "a)". "experience" means "skúsenosti" means "expérience" means "experiencia en el trabajo" means "Erfahrung". "scary" means "strašidelný" means "effrayant" means "Es un susto." means "Angst machen". "by" means "v" means "par" means "por el" means "von". "plane" means "lietadlo" means "plan" means "avión" means "Flugzeug". "have" means "majú" means "ont" means "tiene" means "haben". "before" means "pred" means "avant" means "antes de" means "vor". "not" means "nie" means "pas" means "No" means "nicht". "Travelling" means "Cestovanie" means "Voyage" means "Viajes" means "Reisen". "experienced" means "skúsený" means "expérimenté" means "con experiencia" means "Erfahrene". "of" means "z" means "de" means "de las" means "von". "be" means "byť" means "être" means "ser" means "sein". "can" means "môže" means "peut" means "puede" means "kann". "they've" means "Oni majú" means "Ils ont" means "Ellos han" means "Sie haben". "ages" means "vek" means "âges" means "edad" means "Alter". Translate the following text from English to Slovak:{Source Sentence}
Ours' Prompt	Translate the following sentence from English to Slovak using the provided Dictionary information. Ensure that each dictionary entry is appropriately applied in the translation. For any polysemous words, select the translation that best fits the context of the sentence. Dictionary: {'traumatic': ['traumatické'], 'persons': ['osobám', 'osoby', 'osôb'], 'event': ['podujatie', 'udalosť', 'event', 'udalosti'], 'flown': ['letel', 'leteli'], 'backgrounds': ['pozadia'], 'particularly': ['obzvlášť', 'zvlášť'], 'experience': ['skúsenosť', 'prax', 'skúsenosti'], 'scary': ['strašidelné', 'desivé'], 'plane': ['lietadlo', 'rovina', 'lietadlom'], 'have': ['maj', 'mať', 'majú'], 'before': ['pred', 'predtym', 'predtým'], 'travelling': ['cestovanie'], 'experienced': ['skúsený', 'zažil', 'skúsených', 'skúsenosti', 'skúseného'], 'age': ['age', 'veku', 'vek']}. Sentence: {Source Sentence}
DiPMT's Result (BLEU=9.70)	Cestovanie lietadlom môže byť strašidelná skúsenosť pre ľudí všetkých vekov a pozadiel, najmä ak predtým nesia leteckou cestu alebo majú zažiť traumatičné udalosti.
CoD's Result (BLEU=10.40)	Cestovanie lietadlom môže byť strašidelné skúsenosti pre ľudí všetkých vekov a pozadiel, najmä ak predtým nevyleteli alebo mali traumatuju udalosť.
Ours's Result (BLEU=16.27)	Cestovanie lietadlom môže byť strašidelné zážitok pre osoby všetkých vekov a pozadií, obzvlášť ak predtým neleteli alebo zažili traumtické podujatie.

words (e.g., conjunctions). And both baselines generate ungrammatical outputs (DiPMT: “strašidelná skúsenosť”; CoD: “strašidelné skúsenosti”). Errors stem from mechanical term injection without syntactic/semantic verification. This dilutes focus on semantically critical terms, resulting in grammatical errors and unnatural collocations. In contrast, our constraint-aware prompting filters 54% of low-impact terms and enforces context-aware translation selection, improving BLEU by 67% over CoD. Our Solution’s Efficacy: Keyword filtering reduces prompt terms by 46% (28 → 15), prioritizing high-impact words (e.g., “traumatic”, “scary”). Self-checking iteratively corrects errors. The 67% BLEU improvement over CoD confirms reduced focus dilution.

6. Conclusion

In this paper, we have presented a novel multi-step prompting approach for enhancing the faithfulness and robustness of LLM-based MT. Our method addresses the challenges faced by LLMs in translating rare or specialized terminology by explicitly focusing on key terms in the source sentence and strategically integrating lexical knowledge from high-quality bilingual dictionaries. We further leverage the reflective capabilities of LLMs by employing an iterative self-checking mechanism that allows the model to refine its translations based on both lexical and semantic constraints. Comprehensive experiments conducted on the FLORES-200 benchmark for low-resource languages and contamination-free WMT datasets demonstrate the effectiveness of our approach.

7. Limitation and Future Work

One notable limitation of our approach is its reliance on the quality of the bilingual dictionaries used. The effectiveness of the Retrieval-Augmented Translation and Self-checking methods depends on the accuracy and completeness of these dictionaries. Errors, outdated entries, or incomplete translations in the dictionaries can negatively impact the model’s ability to retrieve accurate word translations, thereby affecting overall translation quality. Additionally, dictionary quality may vary across language pairs and domains, leading to inconsistent performance. To address this, dictionaries should be continuously updated and refined. Future work could focus on dynamically improving dictionary quality or integrating multiple translation data sources to alleviate these issues.

Author Contributions

Shangfeng Chen(332316030987@zzuli.edu.cn): Writing-review & editing, Writing – original draft, Project administration, Data curation, Conceptualization. **Xiayang Shi**(aryang123@163.com): Writing-review & editing, Methodology, Investigation, Funding acquisition. **Pu Li**(superlipu@163.com): Visualization, Supervision, Methodology, Conceptualization. **Jingjing Liu**(2012021@zzuli.edu.cn): Supervision, Methodology, Conceptualization,Funding acquisition. **Yinlin Li**(yinlin.li@ia.ac.cn): Writing–review & editing, Methodology, Conceptualization, Funding acquisition.

Acknowledgements

This work was supported by the Foundation and Cutting-Edge Technologies Research Program of Henan Province (252102211067, 252102210064), the National Natural Science Foundation of China (grant No. 61702516), and the Open Fund of Science and Technology on Thermal Energy and Power Laboratory (No. TPL2020C02), Wuhan 2nd Ship Design and Research Institute, Wuhan, P.R. China, Research and Practice Project on Higher Education Teaching Reform in Henan Province (No. 2024SJGLX0133).

References

[1] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[3] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022.

[4] Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore, December 2023. Association for Computational Linguistics.

[5] Shaolin Zhu, Menglong Cui, and Deyi Xiong. Towards robust in-context learning for machine translation with large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia, May 2024. ELRA and ICCL.

[6] Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. Efficiently exploring large language models for document-level machine translation with in-context learning. *arXiv preprint arXiv:2406.07081*, 2024.

[7] Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. Landermt: Dectecting and routing language-aware neurons for selectively finetuning llms to machine translation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12135–12148. Association for Computational Linguistics, 2024.

[8] Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. *arXiv e-prints*, page arXiv:2302.07856, February 2023.

[9] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics.

[10] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen,

Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics*, 10:50–72, 2022.

[11] Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. Chain-of-dictionary prompting elicits translation in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models. *Trans. Assoc. Comput. Linguistics*, 12:229–246, 2024.

[13] Jonathan Hus and Antonios Anastasopoulos. Back to school: Translation using grammar books. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

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

[16] Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

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

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

[19] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[20] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR, 2023.

[21] Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[22] David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics.

[23] Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[24] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

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

[26] Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. Steering large language models for machine translation with finetuning and in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore, December 2023. Association for Computational Linguistics.

[27] Chunyou Li, Mingtong Liu, Hongxiao Zhang, Yufeng Chen, Jinan Xu, and Ming Zhou. MT2: Towards a multi-task machine translation model with translation-specific in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8616–8627, Singapore, December 2023. Association for Computational Linguistics.

[28] Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada, July 2023. Association for Computational Linguistics.

[29] Jiajun Zhang and Chengqing Zong. Bridging neural machine translation and bilingual dictionaries. *arXiv preprint arXiv:1610.07272*, 2016.

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

[31] Chenggang Mi, Shaolin Zhu, and Rui Nie. Improving loanword identification in low-resource language with data augmentation and multiple feature fusion. *Computational Intelligence and Neuroscience*, 2021(1):9975078, 2021.

[32] Mika Hämäläinen and Khalid Alnajjar. A template based approach for training nmt for low-resource uralic languages - a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 520–525, New York, NY, USA, 2020. Association for Computing Machinery.

[33] Yusen Lin, Jiayong Lin, Shuaicheng Zhang, and Haoying Dai. Bilingual dictionary-based language model pretraining for neural machine translation. *arXiv preprint arXiv:2103.07040*, 2021.

[34] Hao Jiang, Chao Zhang, Zhihui Xin, Xiaoqiao Huang, Chengli Li, and Yonghang Tai. Transfer learning based on lexical constraint mechanism in low-resource machine translation. *Computers and Electrical Engineering*, 100:107856, 2022.

[35] Sicheng Tian, Shaobin Huang, Rongsheng Li, and Chi Wei. A prompt construction method for the reverse dictionary task of large-scale language models. *Engineering Applications of Artificial Intelligence*, 133:108596, 2024.

[36] Sahinur Rahman Laskar, Bishwaraj Paul, Pankaj Dadure, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. English–assamese neural machine translation using prior alignment and pre-trained language model. *Computer Speech & Language*, 82:101524, 2023.

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[38] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[39] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[40] Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. Language models as fact checkers? In Christos Christodoulopoulos, James Thorne, Andreas Vlachos, Oana Cocarascu, and Arpit Mittal, editors, *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online, July 2020. Association for Computational Linguistics.

[41] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[42] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[43] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA, 1998. ACM, Association for Computing Machinery.

[44] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[45] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[46] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[47] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans.*

Assoc. Comput. Linguistics, 10:522–538, 2022.

[48] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

[49] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, 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 2015. Association for Computational Linguistics.

[50] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[51] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[52] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[53] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[54] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study, 2023.