Arabic Dialect Translation with Small LLMs: Enhancing through Reasoning-Oriented Reinforcement Learning

Sohaila Abdulsattar
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
sm10688@nyu.edu

Keith Ross
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
kwr200@nyu.edu

Abstract

Arabic dialect↔English machine translation remains difficult due to extreme dialect variation, inconsistent orthography, and limited parallel data. Moreover, dialect translation is often needed in remote regions or by economically-disadvantaged communities, which often operate in compute-constrained or offline settings. Motivated by these concerns, in this paper we explore optimizing Arabic dialect↔English translators that run over small LLMs, which could be implemented on small offline devices.

We show that reasoning-oriented reinforcement learning can substantially improve small multilingual LLMs for Arabic dialect translation. Using the MADAR corpus, small Qwen-2.5 models trained with a think-then-translate template and optimized with Group-Relative Policy Optimization using a SacreBLEU reward outperform a much larger 7B baseline trained with supervised fine-tuning. The dialect-to-English BLEU score more than doubles from 17.4 to 34.9, while the English-to-dialect COMET score improves from 0.57 to 0.73.

1 Introduction

Machine translation (MT) is now deeply embedded in global communication, supporting access to education, public services, healthcare, and media. Yet, despite recent advances in neural MT, languages with complex internal variation, limited standardized resources, or strong diglossic patterns remain challenging even for the strongest MT models (Tafa et al., 2025; Nicholas and Bhatia, 2023). Arabic is one such prominent example. It is spoken by hundreds of millions across more than twenty countries, but everyday language use is dominated not by a single standard variety, but by a continuum of regional dialects that differ in phonology, morphology, syntax, and lexicon (Zaidan and Callison-Burch, 2014).

These varieties are embedded in a diglossic structure. Modern Standard Arabic (MSA) functions as the formal register used in news, education, and official communication, while regional dialects are the default in speech and much of online interaction. Dialects can diverge to the point of mutual unintelligibility, and they lack both standardized orthography and large-scale parallel corpora (Zaidan and Callison-Burch, 2014). As a consequence, most MT systems for Arabic are trained primarily on MSA, often with limited or noisy dialectal coverage. When exposed to dialectal input, such models frequently misinterpret morphology and dialect-specific vocabulary, leading to mistranslations, semantic drift, and unnatural phrasing (Alhafni et al., 2024).

Recent large language models (LLMs) can often parse dialectal input better than earlier neural MT models, and they achieve competitive scores on benchmarks that mix MSA with a subset of high-resource dialects (Kadaoui et al., 2023). However, the systems that perform best are typically proprietary models with tens or hundreds of billions of parameters, requiring costly infrastructure and stable connectivity. As a result, communities and institutions operating under resource constraints remain least able to benefit from these advances.

This motivates a complementary line of work focused on improving the capabilities of small, open-source models rather than scaling alone (Hsieh et al., 2023; Belcak et al., 2025). We focus on open-source LLMs with fewer than 10 billion parameters. Their compact architectures enable fine-tuning and inference on modest hardware, creating opportunities for localized and offline translation. If we can push compact models closer to the performance of larger systems on challenging tasks like Arabic dialect translation, this would both widen access and challenge the assumption that scaling alone is the dominant path to progress.

In this paper, we explore the use of reasoning-oriented reinforcement learning (RL) for Arabic dialect↔English machine translation in small-scale

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 84–99

large language models. We focus on the MADAR corpus (Bouamor et al., 2018), which provides parallel data for 25 city-level Arabic varieties, MSA, and English, and we use the multilingual Qwen-2.5-base models (1.5B and 3B parameters) as our base architectures (Yang et al., 2025). Starting from baselines using the pretrained Qwen-2.5-7B model and a supervised fine-tuned variant, we then apply Group-Relative Policy Optimization (Shao et al., 2024) to the smaller 1.5B and 3B models, with a think-then-translate output template and with SacreBLEU-based rewards (Post, 2018).

Our work is guided by the following research questions: 1) Can reasoning-oriented reinforcement learning substantially improve Arabic dialect↔English machine translation performance in small multilingual LLMs? 2) How do reinforcement learning-based improvements compare to those obtained through supervised fine-tuning and increased model scale? 3) Does enforcing a think-then-translate output structure in RL lead to the emergence of explicit reasoning behavior during Arabic dialect translation?

Accordingly, the main contributions of our work are as follows:

- We present the first systematic study of reasoning-oriented reinforcement learning for Arabic dialect↔English translation using small multilingual LLMs.
- We show that compact Qwen-2.5 models (1.5B and 3B) optimized with GRPO and a BLEU-based reward substantially outperform a much larger 7B supervised baseline, especially in the dialect→English translation direction.
- We provide a detailed per-dialect evaluation across 25 Arabic city varieties and MSA, highlighting consistent gains even for low-resource and highly divergent dialects.
- We analyze the behavior of the induced reasoning traces and show that translation improvements arise primarily from the models adapting their output strategies to the BLEU-based reward rather than exhibiting more elaborate linguistic reasoning.

2 Related Work

Arabic dialect MT remains difficult due to scarce parallel data, high intra-dialect variability, and the mismatch between informal dialectal text and target language forms (Zbib et al., 2012). Prior research relevant to our work falls into three areas: MT for Arabic dialects, small specialized MT models, and reinforcement learning for inducing structured reasoning in LLMs.

2.1 Arabic Dialects & Machine Translation

Much prior work translates dialects into MSA as a normalization step for downstream Natural Language Processing (NLP) tasks. Shared evaluations such as OSACT 2024 Task 2 (Atwany et al., 2024) specifically assessed dialect→MSA translation across five major varieties (Gulf, Egyptian, Levantine, Iraqi, and Maghrebi). Despite the relative proximity between dialects and MSA, the results remained modest: the top-performing system (based on GPT-3.5) achieved a 29.61 BLEU score, while finetuned encoder-decoder models such as AraT5 and NLLB achieved only 10-12 BLEU scores. A similar trend appears in NADI 2024 Subtask 3 (Abdul-Mageed et al., 2024), another shared task dedicated to dialect→MSA translation; here, systems achieved best scores of around 20 BLEU.

A complementary line of research examines direct dialect↔English translation. Kadaoui et al. (Kadaoui et al., 2023) conducted a comprehensive evaluation in this space by benchmarking ChatGPT, Bard, and Google Translate across ten Arabic dialects. Their results showed modest average scores of 18.2 BLEU for dialect→English and 16.4 BLEU for English→dialect, with substantial variation across dialects. The authors concluded that although LLMs outperform standard neural MT baselines, they still struggle to capture the idiomatic, cultural, and pragmatic nuances essential to high-quality dialect translation.

These observations are further supported by AraDiCE (Mousi et al., 2025), a large-scale benchmark designed to assess both dialectal understanding and cultural competence in LLMs. In its translation subset, even Arabic-centric models such as Jais-13B and AceGPT-13B achieved only 13-17 BLEU scores for dialect→English and 8-11 for English→dialect. Together, these evaluations also indicate a consistent pattern: current LLMs tend to understand dialectal input reasonably well but they underperform in generation, thus often performing better in translating *from* than *into* Arabic dialects.

2.2 Small, Specialized MT Models

Recent work has shown that small, task-specialized MT models can rival or outperform much larger architectures when trained on high-quality or domain-specific data. This is especially important for dialectal Arabic, where data is scarce and deployability matters.

SMaLL-100 (Mohammadshahi et al., 2022) distills M2M-100 12B into 200-600M parameter models that remain competitive on many (non-Arabic) low-resource pairs. Mutarjim (Hennara et al., 2025), a 1.5B Arabic↔English model, reaches 61.4 ChrF++ and 0.83 COMET on the Tarjama-25 Arabic benchmark and surpasses proprietary models far larger in scale like GPT-4o mini.

However, none of these works consider Arabic dialect translation, and unlike prior efforts that rely on distillation or supervised fine-tuning, we investigate whether reinforcement learning can push small models further by improving their ability to plan, reason, and generate coherent translations.

2.3 Reinforcement Learning for Reasoning and Machine Translation

Early progress on LLM reasoning came from prompting, with chain of thought methods (CoT) (Wei et al., 2022) improving reasoning without altering model weights. Reinforcement learning (RL) approaches then aimed to teach explicit reasoning behaviors in LLMs. Shao et al. (2024) proposed Group-Relative Policy Optimization (GRPO), an RL method designed specifically for LLMs that avoids value modeling by comparing outputs in grouped batches. Utilizing GRPO, Guo et al. (2025) introduced DeepSeek-R1-Zero, which demonstrated that structured reasoning can emerge in LLMs from reward-only training.

Directly connecting RL-based reasoning to MT, He et al. (2025) introduced a reason-then-translate framework where models produce structured thinking steps before generating the translation. Their RL procedure (based on CoT templates and two-stage optimization) yielded improvements in both accuracy and fluency. Recent work by Feng et al. (2025) produced the MT-R1-Zero model, which adapts DeepSeek-R1-Zero's RL framework (Guo et al., 2025) specifically for machine translation. Instead of producing a translation directly, their model is trained to output a reasoning step before the final translation, with reward functions that evaluate both the reasoning format and the final translation. Their results show that reasoning emerges purely from reward design and that it improves translation quality even without supervised CoT data.

Together, this research establishes the groundwork for our approach while highlighting a clear gap: small multilingual LLMs have not been systematically explored for Arabic dialect translation, and reasoning-oriented reinforcement learning has not been applied in this setting. We address this gap by testing whether RL-driven reasoning signals can enhance dialect↔English translation in compact models and reduce their performance gap with larger systems.

Code	City	Code	City
ALE	Aleppo	ALG	Algiers
ALX	Alexandria	AMM	Amman
ASW	Aswan	BAG	Baghdad
BAS	Basra	BEI	Beirut
BEN	Benghazi	CAI	Cairo
DAM	Damascus	DOH	Doha
FES	Fes	JED	Jeddah
JER	Jerusalem	KHA	Khartoum
MOS	Mosul	MUS	Muscat
RAB	Rabat	RIY	Riyadh
SAL	Salt	SAN	Sana'a
SFX	Sfax	TRI	Tripoli
TUN	Tunis		

Table 1: Mapping between the MADAR corpus dialect codes and their corresponding cities.

3 Methodology

3.1 Data and Preprocessing

We use the MADAR corpus (Bouamor et al., 2018), a large-scale resource for Arabic dialect machine translation created as part of the MADAR Project. The corpus is based on English sentences from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2007), which were translated into 25 Arabic city dialects, as well as MSA.

Since MADAR is manually translated, aligned, and quality-controlled, no additional cleaning or filtering was required beyond tokenization. We use the MADAR-26 configuration of the MADAR corpus, which consists of 2,000 aligned sentence pairs per dialect and MSA, for a total of 52,000 pairs. We refer to dialectal varieties using the corpus's standard short codes; Table 1 provides a mapping between the codes and their corresponding city dialects.

3.2 Baseline Models

To study the impact of reasoning-oriented RL on Arabic dialect translation, we use the Qwen-2.5-base family of multilingual LLMs (Yang et al., 2025). Qwen-2.5 models are decoder-only transformers with strong multilingual performance and substantial Arabic training data, making them suitable for dialectal translation. Prior work by Feng et al. (2025) also shows that Qwen-2.5 exhibits stable and faithful reasoning behavior under reinforcement learning, rather than attempting to "hack" the reward structure. For these reasons, we adopt Qwen-2.5 as the base model family for all experiments.

Although our RL-trained models use the smaller 1.5B and 3B variants, we use the Qwen-2.5-7B model as an upper-bound baseline. Its larger scale provides a meaningful reference point: if RL-trained smaller models approach or exceed its performance, this suggests gains beyond scale alone. We evaluate the 7B model both from its pretrained checkpoint directly and after supervised fine-tuning on MADAR-26 using standard next-token prediction. Together, these baselines capture the performance achievable through conventional training and allow us to isolate the effects of reasoning-oriented reinforcement learning. For completeness, we also report baseline performance for the 1.5B and 3B models in Appendix Tables 5, 6, 8, and 9.

For each dialect in MADAR-26, we adopt a simple, direct translation prompt: "Translate the following {DialectCity} Arabic text into English:" or "Translate the following English text into {DialectCity} Arabic:", followed by the source sentence. This prompt design is deliberately minimal, as recent findings by Zheng et al. (2024) show that simple instruction-style translation prompts outperform more elaborate or template-heavy formats.

3.2.1 Pretrained baselines

As a first baseline, we evaluate the pretrained Qwen-2.5-7B model out-of-the-box, without any task-specific fine-tuning or adapters. We load Qwen-2.5-7B in evaluation mode and generate translations deterministically using greedy decoding with a maximum generation length of 64 tokens and a mild repetition penalty of 1.1. Prompts are tokenized with the native Qwen tokenizer, padded and truncated to a maximum length of 512 tokens, and fed to the model in mini-batches of size eight. After generation, we slice off the prompt portion of the output to isolate only the model-produced continuation and decode it into plain text.

3.2.2 Supervised fine-tuning baselines

To establish stronger baselines, we fine-tune Qwen-2.5-7B using the same instruction-style prompts mentioned above. We frame translation as a causal language modeling problem: the model is given a natural-language instruction and the source sentence as context, and is trained to generate the target sentence as a continuation. Additional fine-tuning details are provided in Appendix A.1.

We train a multi-dialectal model with an equal representation for all 26 dialect varieties in the training set. To make the model explicitly aware of the source variety when translating from Arabic dialects, we add a lightweight control token of the form <SRC:DIALECT> to the tokenizer (e.g., <SRC:CAI>, <SRC:ALG>). This token is included in the prompt but masked during loss computation. This setup encourages the model to learn both common cross-dialect structure and the fine-grained distinctions across regional varieties.

3.3 Reinforcement Learning Framework

We next fine-tune Qwen-2.5-based policies with a group-based policy-gradient method and a reward signal that combines (1) a strict constraint on output format and (2) a lexical MT quality metric. This setup follows the general MT-R1-Zero framework proposed by Feng et al. (2025) but instantiated for Arabic dialect translation.

3.3.1 Group-Relative Policy Optimization

For RL, we adopt GRPO, introduced in the DeepSeek (Shao et al., 2024), which is a simplified version of Proximal Policy Optimization (PPO). GRPO removes the need for the learned critic in PPO and has been shown to yield stable optimization for small and medium-sized LLMs.

During training, for each translation query q, we sample a group of G candidate outputs $\{o_1, o_2, \ldots, o_G\}$ from the frozen sampling policy $\pi_{\theta_{\text{old}}}$. Each output o_i receives a scalar reward r_i computed using a rule-metric mixed reward (described in the next section). GRPO computes an advantage for each sample via:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G) + \epsilon}, \qquad (1)$$

where A_i is the group-normalized advantage and ϵ is a small constant to stabilize variance.

The policy is then updated by maximizing the clipped GRPO objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o_i\sim\pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \min\Big(\rho_i A_i, \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)\, A_i\Big)\Bigg] - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \tag{2}$$

where the likelihood ratio is

$$\rho_i = \frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}. \tag{3}$$

The Kullback-Leibler (KL) divergence penalty is included to prevent the updated policy from drifting too far from a fixed reference model π_{ref} (the initialized model). The coefficient β controls the strength of this constraint, while ε determines the PPO clipping range.

3.3.2 Reasoning-structured output format

Following prior reasoning-oriented reinforcement learning setups (Guo et al., 2025), we require the model to separate its internal reasoning from its final translation output using a fixed template, as their results indicate that pure reinforcement learning with strict formatting can elicit intermediate reasoning without gold reasoning traces.

For each input sentence, the model is prompted to think step-by-step and then produce the translation in the following structured format:

```
<think> reasoning process </think>
<translate> final output </translate>
```

We use the same prompting scheme as MT-R1-Zero (Feng et al., 2025), reproduced in Appendix B, in order to isolate the effect of reinforcement learning in our setting. The content inside `<think>` is treated as a latent reasoning trace and is discarded at evaluation time; only the span inside `<translate>` is used as the system's translation.

3.3.3 Reward Design

In line with Feng et al. (2025), we employ a rule-metric mixed reward that combines a format reward and a metric reward.

We first check whether o (a sampled model output) conforms to the required structure, with no malformed or repeated tags. The format reward is defined as:

$$S_{\text{format}}(o) = \begin{cases} 1, & \text{if output format is correct,} \\ -1, & \text{otherwise} \end{cases} \tag{4}$$

If the format is incorrect, we do not compute translation quality and instead assign a fixed penalty. This encourages the model to first learn to reliably respect the reasoning template.

When the output is correctly formatted, we use the BLEU score as our MT metric. Given a generated translation *trans* and a reference *ref*, the metric reward is:

$$S_{\text{metric}}(o) = B(\text{trans, ref}) \tag{5}$$

where $B(\cdot,\cdot)$ denotes the normalized BLEU score computed via SacreBLEU (Post, 2018). We deliberately choose BLEU over semantic metrics to reduce computational overhead during RL and to provide clearer token-level learning signals for the dialectal pairs.

The final scalar reward $r(o)$ combines the two components following the rule-based structure explained above:

$$r(o) = \begin{cases} S_{\text{format}}(o) - 2, \\ \quad \text{if } S_{\text{format}}(o) = -1, \\ S_{\text{format}}(o) + S_{\text{metric}}(o), \\ \quad \text{if } S_{\text{format}}(o) = 1. \end{cases} \tag{6}$$

Thus, misformatted outputs receive a fixed negative reward, while correctly formatted outputs receive a continuous reward in the range $[1, 2]$, depending on BLEU quality. This design provides both a hard constraint on structure and fine-grained feedback on translation quality, and can be plugged directly into the GRPO objective described above.

3.3.4 Experimental Setup

We apply GRPO to two model sizes from the Qwen-2.5-base family (Yang et al., 2025): a 1.5B and a 3B parameter model. For each size, we train two systems: dialect→English and English→dialect.

For the dialect→English direction, we initialize RL directly from the pretrained Qwen-2.5 checkpoint. However, for the inverse English→dialect direction, this initialization resulted in very weak dialectal generation. We, therefore, introduced a supervised fine-tuning warm up on the MADAR-26 sentence pairs, followed by RL from the resulting checkpoints. The supervised stage serves only to stabilize dialect generation before reasoning is shaped through RL.

Full implementation and training details are provided in Appendix A.2 and in our public training scripts.[1]

4 Evaluation

4.1 Evaluation Metrics

We evaluate all models using SacreBLEU (Post, 2018) and COMET-DA (Rei et al., 2022), which capture complementary aspects of translation quality. SacreBLEU measures n-gram overlap between system outputs and reference translations, providing a standardized assessment of lexical fidelity and word-level alignment. While sensitive to morphological and orthographic variation, it remains useful for identifying dialect-specific lexical errors.

To complement this, we use the COMET-DA model, which predicts semantic adequacy by jointly encoding the source, hypothesis, and reference. Unlike BLEU, COMET rewards meaning preservation even when lexical realization differs from the reference. Using both metrics allows us to disentangle lexical fidelity from semantic adequacy, offering a more comprehensive evaluation of model behavior across the diverse set of Arabic dialects considered in our work.

4.2 Evaluation Overview

Table 2 summarizes the translation performance across all training approaches and both translation directions; these are the overall scores computed over the entire test set consisting of all 25 dialects and MSA. Figure 1 provides a visual summary of these overall BLEU and COMET scores across models. Per-dialect scores, obtained by filtering test instances by dialect, are also reported in the Appendix: Tables 7 and 10 for the baseline models, and Tables 11 and 12 for the RL-trained models.

In what follows, we highlight a small number of representative scores to illustrate systematic trends, focusing on how model behavior evolves across training approaches.

4.3 Baseline Models

4.3.1 Pretrained Performance

Without fine-tuning, Qwen-2.5-7B demonstrates partial semantic understanding of dialectal Arabic, but little control over dialectal generation.

[1]https://github.com/Sohaila-Abdulsattar-Mohammed/Arabic-Dialect-Translation-with-Small-LLMs-Enhancing-through-Reasoning-Oriented-RL

	BLEU	COMET
Dia→En		
Pretrained (7B)	7.60	0.58
SFT (7B)	17.37	0.61
RL (1.5B)	27.63	0.72
RL (3B)	**34.88**	**0.79**
En→Dia		
Pretrained (7B)	0.56	0.46
SFT (7B)	11.10	0.57
RL (1.5B)	10.70	0.71
RL (3B)	**11.31**	**0.73**

Table 2: Performance comparison of Qwen-2.5 across training settings for Dialect→English and English→Dialect translation. Best results within each direction are bolded.

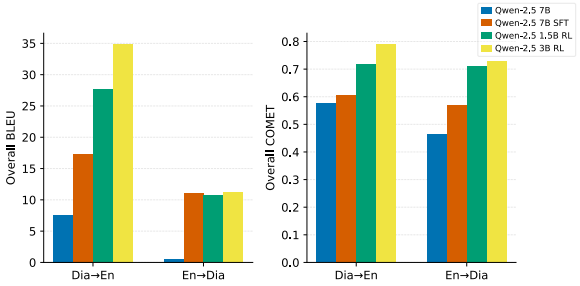

Figure 1: Overall BLEU and COMET scores on the multidialect MADAR-26 test set for all evaluated models. Our RL-trained models substantially outperform both the pretrained and SFT baselines, with the largest gains appearing in the Dialect→English direction.

In the Dialect→English direction, the model achieves 7.6 BLEU and 0.58 COMET overall (Table 2). However, Appendix Table 7 shows that the overall score masks wide variation. Stronger dialects such as MSA (19.05 BLEU) and MUS (14.16) contrast sharply with the weakest dialects, particularly SFX (3.04), TUN (3.20), and RAB (4.15). The North African dialects yielding the weakest scores is consistent with their greater divergence from MSA and with their relatively low representation in multilingual pretraining corpora (Kwaik et al., 2018).

The English→Dialect direction fails more dramatically. Overall BLEU drops to 0.56, with most dialects scoring below 0.7 BLEU. Even the strongest dialect, MSA, reaches only 2.52 BLEU, while several dialects are effectively non-functional with near-zero scores. However, COMET scores remain relatively moderate compared to BLEU (0.46 overall) because the model often defaults to MSA instead of producing dialectal forms, as illustrated

Source (EN)	We'd like to stay for four nights from August first.
Reference (ASW)[2]	إحنا كنا عاوزين نقعد أربع ليالي من أول يوم واحد في شهر أغسطس.
Model Final Translation[3]	نريد أن نقيم لمدة أربعة ليالٍ من تاريخ الأولّ من أغسطس.

Table 3: Qualitative example from the pretrained Qwen-2.5-7B evaluation. The model produces an MSA-style translation instead of the target dialectal form.

in Table 3. As a semantic metric, COMET assigns non-trivial scores to these outputs despite their stylistic mismatch with the dialectal references.

Notably, zero-shot scores reflect not only weak translation quality but also frequent violations of the required output format, with the model often generating extraneous text instead of a direct translation.

4.3.2 Supervised Fine-Tuning Performance

Supervised fine-tuning substantially reshapes the pretrained behavior, with both BLEU and COMET scores increasing markedly across the board.

For Dialect→English, SFT more than doubles overall BLEU from 7.6 to 17.37 and raises COMET from 0.58 to 0.61 (Table 2). Crucially, the weakest dialects show substantial absolute gains under SFT: SFX improves from 3.04 to 12.92 BLEU, TUN from 3.20 to 13.67, and RAB from 4.15 to 15.96 (Appendix Table 10). This indicates that SFT does not merely amplify already-strong dialects, but teaches the model to systematically interpret dialectal variation.

For English→Dialect, Overall BLEU increases from near-zero to 11.10, showing that the model begins to generate dialectal outputs instead of defaulting to MSA. However, COMET remains noticeably higher (0.57 overall), reflecting a persistent gap between semantic adequacy and surface-level dialectal accuracy. This suggests that while supervised fine-tuning enables dialectal generation, the produced outputs are often only approximate realizations of the target dialect.

[2]Transliteration: *ihna kuna 'awzin nu'ud arba' layali min awwal yom wahid fi shahr aghustus*

[3]*nurid an nuqim li-muddat arba' layalin min tarikh al-awwal min aghustus.*

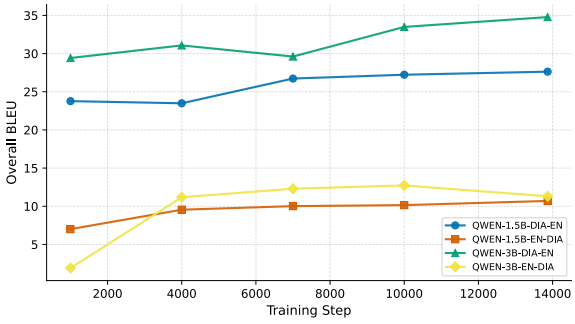

Figure 2: BLEU progression during GRPO training for all four RL models. Dialect→English models start from the pretrained checkpoints, while English→Dialect models begin from their corresponding SFT warm-up checkpoints.

4.4 Reinforcement Learning Models

Figure 2 summarizes BLEU progression throughout RL training. RL yields the largest gains overall, but its effect differs markedly by translation direction.

In Dialect→English, RL produces consistent and substantial improvements across all dialects. The 1.5B model reaches 27.63 BLEU, while the 3B model attains 34.88 BLEU overall (Table 2). Appendix Table 12 shows that even previously weak dialects benefit significantly: SFX improves from 12.92 (SFT) to 20.77 BLEU under 3B RL, TUN from 13.67 to 23.00, and RAB from 15.96 to 26.96. At the same time, strong dialects such as MSA (45.06 BLEU) and RIY (44.07) also improve, indicating that RL does not merely rebalance performance but raises the ceiling across the board.

Scaling from 1.5B to 3B yields uniform gains, typically in the 7 – 10 BLEU range per dialect, with no observed regressions. COMET scores similarly concentrate in the high 0.8 range for most dialects, signaling improved semantic consistency.

In English→Dialect, we first experimented with applying GRPO directly to the pretrained Qwen-2.5 1.5B and 3B models. In this configuration, the 1.5B model reaches an overall BLEU of only 1.46 and COMET of 0.53, while the 3B model reaches 2.01 BLEU and 0.58 COMET (Appendix Table 13). To address these weak results, we first fine-tuned the models in the English→dialect direction, then initialized RL training from these SFT checkpoints. This SFT+RL training setup leads

to much stronger models. While BLEU remains similar to SFT (10.70 for 1.5B and 11.31 for 3B), COMET improves substantially from 0.57 (SFT) to 0.71-0.73. For example, RIY improves from 13.19→13.69 BLEU but from 0.53→0.79 COMET, suggesting that RL primarily enhances semantic adequacy and preference alignment rather than exact surface matching when generating dialectal Arabic.

4.4.1 Analysis of Reasoning Behavior

A central motivation of our reinforcement learning setup was to encourage the emergence of explicit reasoning behavior during translation, following a reasoning-first paradigm similar in spirit to R1-zero approaches (Guo et al., 2025). To this end, we structured the model output to separate a `<think>` phase from the final `<translate>` output, and optimized the model using pure reinforcement learning with a BLEU-based reward, without any supervised reasoning traces.

Despite this design, our models trained with RL do not exhibit meaningful or structured reasoning behavior. Across both translation directions and model sizes, the `<think>` segments rarely contain step-by-step linguistic analysis, such as decomposition of dialectal morphology, lexical disambiguation, or syntactic planning. Instead, the reasoning traces frequently collapse into superficial or degenerate patterns. Common failure modes include generic placeholder statements, restatements of the input prompt, or direct duplication of the final translation within the `<think>` block. Representative examples of this behavior are shown in Table 4.

Importantly, this absence of explicit reasoning persists even as overall translation quality improves substantially under RL. This indicates that the gains achieved through RL are not driven by improved reasoning quality, but rather by the direct optimization of the BLEU-based reward. In practice, the model learns to produce translations that score well under the reward function without internalizing or externalizing a structured reasoning process.

Source (ALE)[4]	وين بوابة الصعود ع هالطيارة؟
Reference (EN)	Where is the boarding gate for this flight?
Model Full Output	<think> The Aleppo Arabic sentence </think> <translate>Where is the boarding gate for this flight? </translate>
Source (EN)	I'd like to rent a car of this class for two days.
Reference (CAI)[5]	عايز أأجر عربية من الفئة ديه لمده يومين.
Model Full Output[6]	<think>عايز أأجر سيارة من هذا الفئة لمدة يومين. </think> <translate>عايز أأجر سيارة من الفئة دي لمدة يومين. </translate>

Table 4: Examples of `<think>` segments produced by the Qwen-2.5-3B RL-trained models, illustrating that the reasoning segments typically collapse into shallow or generic content rather than meaningful step-by-step analysis.

5 Conclusion

Our work tackled whether reasoning-oriented reinforcement learning can make small multilingual LLMs competitive for Arabic dialect↔English translation, a setting that is both linguistically challenging and low-resource. Using the MADAR corpus as a multidialect testbed and Qwen-2.5 1.5B and 3B models as base architectures, we adapted the R1-Zero RL framework to a multi-dialect setting with a think-then-translate template, SacreBLEU-based rewards, and SFT warmup where needed. Overall, our GRPO-trained models consistently and substantially improved over both the pretrained and SFT Qwen-2.5-7B baselines, particularly in the dialect→English direction, and achieved strong COMET gains for English→dialect.

Our results offer two main takeaways. First, from a practical perspective, they demonstrate that small, open-source models when aligned with carefully designed reward signals can outperform larger systems on challenging, low-resource MT tasks, while remaining deployable on modest hardware. This challenges the assumption that scale is the primary route to better MT and suggests that RL can be an effective lever for unlocking the potential of compact architectures, especially in settings where cost, la-

[4] Transliteration: *wayn bawwabet al-su'ud 'a hal-tayyara?*
[5] *'ayiz a'aggar 'arabiyya min al-fi'a dih li-muddit yomein.*
[6] *'ayiz a'aggar sayyara min hadha al-fi'a li-muddat yomein; 'ayiz a'aggar 'arabiyya min al-fi'a dih li-muddit yomein.*

tency, and offline operation matter. Second, from a methodological perspective, our experiments show that RL optimized with a BLEU-based lexical reward and a structured output format is sufficient to yield substantial improvements in lexical fidelity and semantic adequacy. Notably, these gains are achieved even though the intermediate `<think>` segments, which were intended to encourage deliberation, do not exhibit meaningful emergent reasoning in our setting.

Finally, by showing that small RL-aligned models can close and, in some cases, surpass the gap with larger systems on Arabic dialect translation, we hope to encourage the MT and Arabic NLP communities to invest further in open, deployable models tailored to dialectal realities. Extending these methods to richer dialectal corpora, additional Arabic varieties, and other diglossic or low-resource language families could broaden access to high-quality MT for communities that are currently underserved by large proprietary systems.

6 Limitations

Our results should be interpreted in light of several limitations related to data scope, task difficulty, and training design.

All experiments are conducted on the MADAR corpus, which is restricted to sentence-level translation in the travel domain; consequently, the observed gains may not generalize to longer contexts, more diverse genres, or naturally occurring code-switching, which remain underrepresented in available dialectal resources.

English→Dialect translation remains substantially more challenging than Dialect→English. While RL improves semantic adequacy as reflected by COMET, BLEU scores remain modest. This reflects the inherent complexity of modeling non-standardized dialectal varieties and their fine-grained surface realizations. Our results therefore highlight both the promise of RL for improving small models' dialectal generation and how there remains considerable headroom for future work in this area.

Finally, while our approach is motivated by *reasoning-oriented* reinforcement learning, we study it under a controlled instantiation involving small models and a fixed reward design. This allows us to isolate the effects of reasoning-oriented RL but does not explore the broader design space of mechanisms that may further encourage explicit reasoning behavior in translation.

Taken together, these limitations highlight how future work could explore richer reward formulations, longer-context and multi-domain evaluation, and alternative training regimes that more directly target semantic robustness, dialectal fidelity, or interpretable reasoning behavior.

7 Acknowledgments

This work is submitted in part by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the Research Institute Award CG010.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. OSACT 2024 task 2: Arabic dialect to MSA translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98–103, Torino, Italia. ELRA and ICCL.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and 1 others. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, and 1 others. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*.

Khalil Hennara, Muhammad Hreden, Mohamed Motaism Hamed, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Mutarjim: Advancing bidirectional arabic-english translation with a small language model. *arXiv preprint arXiv:2505.17894*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El-Moatez-Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia Computer Science*, 142:2–13. Arabic Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. Aradice: Benchmarks for dialectal and cultural capabilities in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

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

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. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Taofik O Tafa, Siti Zaiton Mohd Hashim, Mohd Shahizan Othman, Hitham Alhussian, Maged Nasser, Said Jadid Abdulkadir, Sharin Hazlin Huspi, Sarafa O Adeyemo, and Yunusa Adamu Bena. 2025. Machine translation performance for lowresource languages: A systematic literature review. *IEEE Access*.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *Preprint*, arXiv:2402.15061.

A Training Details

A.1 Supervised Fine-tuning

We fine-tune Qwen-2.5 models using parameter-efficient LoRA adapters applied to all attention and MLP projection layers (`q,k,v,o,up,down,gate`), with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Training is performed for two epochs with an effective batch size of 8 (per-device batch size 1 with gradient accumulation), a learning rate of 5×10^{-5}, and a maximum sequence length of 512 tokens. Mixed-precision training is used, with gradient checkpointing enabled to reduce memory usage.

All experiments are run on a single node with two NVIDIA A100 GPUs (80GB), using distributed data-parallel training.

A.2 Reinforcement Learning

All reinforcement learning experiments are conducted using the `verl`[7] framework with GRPO as the advantage estimator.

We train reinforcement learning models using GRPO on Qwen-2.5 models with 1.5B and 3B parameters, initialized either from the pretrained checkpoints (Dialect→English) or from the corresponding SFT checkpoints (English→Dialect).[8] Training is performed for one epoch with a per-step batch size of 3 and three rollouts per prompt. The actor learning rate is set to 5×10^{-7}. Maximum prompt and response lengths are set to 384 and 768 tokens, respectively. Mixed-precision training is used, with gradient checkpointing and fully sharded data parallelism enabled to reduce memory usage.

For each group of rollouts generated by the frozen sampling policy, we perform a single GRPO update before refreshing the policy. We use a PPO clipping parameter of $\varepsilon = 0.2$, while no explicit KL regularization is applied during training ($\beta = 0.0$).

All reinforcement learning experiments are run on a single node with three NVIDIA A100 GPUs (80GB).

B Reinforcement Learning Training Prompt

For all RL experiments, we use the same prompting scheme as MT-R1-Zero (Feng et al., 2025). We reproduce it here for completeness and reproducibility:

> A conversation between User and Assistant. The User asks for a translation from {source language/dialect name} to {target language/dialect name}, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the user with the final translation. The reasoning process and final translation are enclosed within <think> </think> and <translate> </translate> tags, respectively, i.e., <think> reasoning process here </think><translate> final translation here </translate>.
>
> User:{source to be translated}
> Assistant:

[7] https://github.com/volcengine/verl

[8] The SFT checkpoints used to initialize English→Dialect reinforcement learning are trained using the same procedure described in Section A.1.

C Per-Dialect Evaluation Results of the Qwen-2.5-1.5B Pretrained Baseline

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	3.41	0.50	0.23	0.40
ALE	2.60	0.47	0.16	0.40
ALG	2.68	0.51	0.10	0.42
ALX	4.11	0.49	0.11	0.40
AMM	3.38	0.51	0.17	**0.44**
ASW	3.76	**0.53**	0.21	0.42
BAG	3.22	0.48	0.15	0.40
BAS	2.87	0.50	0.13	**0.44**
BEI	2.21	0.45	0.10	0.39
BEN	3.14	0.50	0.28	0.40
CAI	3.94	0.50	0.19	0.39
DAM	2.52	0.46	0.13	0.38
DOH	3.65	0.50	0.25	0.40
FES	3.59	0.50	0.30	0.41
JED	2.84	0.50	0.24	0.39
JER	2.49	0.46	0.10	0.37
KHA	**5.91**	**0.56**	0.22	**0.46**
MOS	2.60	0.46	0.14	0.41
MSA	**7.99**	**0.59**	**0.88**	0.38
MUS	**5.95**	0.53	**0.62**	0.38
RAB	2.14	0.48	0.12	0.37
RIY	4.69	0.52	**0.35**	0.40
SAL	4.03	0.49	0.08	0.38
SAN	3.22	0.50	0.14	0.41
SFX	1.81	0.44	0.22	0.37
TRI	2.54	0.50	0.09	0.40
TUN	1.87	0.44	0.09	0.37

Table 5: Per-dialect evaluation results for pretrained Qwen-2.5-1.5B (out-of-the-box) across all MADAR-26 dialects. Highlighted values indicate the top three scores within each metric column.

D Per-Dialect Evaluation Results of the Qwen-2.5-3B Pretrained Baseline

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	3.94	0.51	0.29	0.42
ALE	3.44	0.49	0.19	0.38
ALG	2.95	0.53	0.47	0.52
ALX	4.25	0.52	0.15	0.37
AMM	4.96	0.53	0.47	**0.59**
ASW	**5.96**	**0.57**	0.50	**0.60**
BAG	4.20	0.52	0.06	0.38
BAS	4.01	0.53	0.24	0.56
BEI	3.82	0.46	0.12	0.36
BEN	2.89	0.50	0.27	0.38
CAI	3.46	0.51	0.20	0.36
DAM	3.48	0.51	0.12	0.37
DOH	3.41	0.51	0.28	0.44
FES	3.83	0.50	0.19	0.37
JED	4.04	0.53	0.25	0.42
JER	3.62	0.50	0.15	0.37
KHA	**8.06**	**0.60**	**1.35**	**0.64**
MOS	4.25	0.49	0.13	0.36
MSA	**8.64**	**0.63**	**0.89**	0.38
MUS	4.96	0.55	**0.56**	0.37
RAB	2.69	0.51	0.14	0.37
RIY	5.01	0.54	0.23	0.40
SAL	3.40	0.49	0.14	0.37
SAN	3.48	0.50	0.38	0.57
SFX	1.27	0.44	0.14	0.36
TRI	3.11	0.50	0.12	0.38
TUN	1.72	0.43	0.10	0.36

Table 6: Per-dialect evaluation results for pretrained Qwen-2.5-3B (out-of-the-box) across all MADAR-26 dialects. Highlighted values indicate the top three scores within each metric column.

E Per-Dialect Evaluation Results of the Qwen-2.5-7B Pretrained Baseline

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	7.60	0.58	0.56	0.46
ALE	8.25	0.59	0.61	0.45
ALG	7.25	0.60	0.27	0.51
ALX	8.40	0.56	0.39	0.47
AMM	8.38	0.55	0.47	0.54
ASW	7.77	0.61	0.65	**0.59**
BAG	7.69	0.58	0.27	0.43
BAS	7.05	0.59	0.33	0.56
BEI	4.97	0.49	0.15	0.43
BEN	6.05	0.52	0.42	0.44
CAI	7.83	0.57	0.41	0.45
DAM	6.66	0.51	0.42	0.47
DOH	8.21	0.61	0.53	0.48
FES	6.89	0.55	0.57	0.40
JED	8.40	0.60	0.48	0.48
JER	8.13	0.59	0.33	0.42
KHA	**13.84**	**0.64**	0.79	**0.57**
MOS	6.24	0.52	0.25	0.42
MSA	**19.05**	**0.74**	**2.52**	0.50
MUS	**14.16**	**0.67**	**1.33**	0.41
RAB	4.15	0.54	0.17	0.40
RIY	7.78	0.53	**0.93**	0.47
SAL	10.07	0.61	0.43	0.42
SAN	8.15	0.62	0.57	**0.58**
SFX	3.04	0.50	0.26	0.40
TRI	7.47	0.59	0.24	0.43
TUN	3.20	0.46	0.09	0.36

Table 7: Per-dialect evaluation results for pretrained Qwen-2.5-7B (out-of-the-box) across all MADAR-26 dialects. Highlighted values indicate the top three scores within each metric column.

F Per-Dialect Evaluation Results of the Qwen-2.5-1.5B SFT Baseline

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	11.11	0.55	8.69	0.60
ALE	9.53	0.53	9.88	0.58
ALG	10.12	0.56	7.57	0.61
ALX	12.68	0.56	7.42	0.58
AMM	9.79	0.56	**14.07**	**0.72**
ASW	11.14	**0.59**	12.06	**0.72**
BAG	11.72	0.56	5.78	0.59
BAS	10.80	0.54	10.77	0.69
BEI	11.23	0.53	4.83	0.55
BEN	11.24	0.56	**12.55**	0.61
CAI	9.90	0.56	4.72	0.54
DAM	12.12	0.54	6.19	0.57
DOH	9.79	0.54	12.06	0.60
FES	11.29	0.55	6.54	0.53
JED	10.14	0.55	8.77	0.62
JER	11.15	0.53	9.08	0.58
KHA	**13.47**	**0.61**	**14.46**	**0.75**
MOS	10.96	0.52	8.07	0.60
MSA	**12.84**	**0.60**	10.47	0.54
MUS	11.82	0.54	8.82	0.59
RAB	9.15	0.54	2.97	0.51
RIY	11.81	0.56	12.12	0.63
SAL	10.13	0.53	11.05	0.59
SAN	**13.74**	0.57	7.04	0.69
SFX	8.47	0.50	2.22	0.51
TRI	11.93	0.56	6.63	0.57
TUN	7.28	0.53	2.88	0.49

Table 8: Supervised fine-tuning per-dialect evaluation results for Qwen-2.5-1.5B across all MADAR-26 dialects. Bolded values indicate the top three scores within each metric column.

G Per-Dialect Evaluation Results of the Qwen-2.5-3B SFT Baseline

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	10.24	0.53	8.82	0.57
ALE	9.73	0.50	12.06	0.57
ALG	10.52	0.54	8.80	0.62
ALX	11.20	0.53	11.23	0.57
AMM	10.27	0.54	**16.22**	**0.73**
ASW	11.98	**0.58**	11.94	**0.75**
BAG	8.46	0.52	2.88	0.46
BAS	9.28	0.51	10.15	0.68
BEI	9.87	0.51	4.57	0.55
BEN	8.63	0.52	12.87	0.58
CAI	9.40	0.54	6.81	0.50
DAM	9.25	0.50	8.35	0.56
DOH	10.49	0.50	11.43	0.58
FES	**13.75**	0.54	5.58	0.49
JED	10.32	0.53	7.13	0.58
JER	10.04	0.50	**13.06**	0.57
KHA	**15.77**	**0.61**	**17.48**	**0.78**
MOS	6.95	0.50	8.31	0.53
MSA	9.30	**0.55**	3.83	0.38
MUS	9.31	0.51	5.47	0.51
RAB	10.40	0.55	3.89	0.49
RIY	8.66	0.51	8.23	0.48
SAL	8.31	0.50	10.81	0.52
SAN	**12.69**	0.55	10.17	0.71
SFX	9.04	0.51	1.35	0.47
TRI	11.81	0.55	7.25	0.57
TUN	9.61	0.51	3.88	0.52

Table 9: Supervised fine-tuning per-dialect evaluation results for Qwen-2.5-3B across all MADAR-26 dialects. Bolded values indicate the top three scores within each metric column.

H Per-Dialect Evaluation Results of the Qwen-2.5-7B SFT Baseline

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	17.37	0.61	11.10	0.57
ALE	18.58	0.59	**16.82**	0.60
ALG	14.56	0.61	12.27	0.66
ALX	**20.62**	0.62	7.30	0.46
AMM	17.74	0.62	**18.97**	**0.74**
ASW	19.27	0.64	13.52	**0.76**
BAG	17.73	0.60	7.97	0.54
BAS	15.92	0.60	15.54	0.70
BEI	12.59	0.57	6.66	0.51
BEN	18.58	0.61	13.55	0.51
CAI	17.60	0.62	5.26	0.44
DAM	16.69	0.60	10.64	0.52
DOH	17.94	0.60	13.09	0.56
FES	18.31	0.62	9.71	0.53
JED	14.71	0.60	9.22	0.55
JER	16.08	0.58	10.62	0.50
KHA	**21.96**	**0.68**	**21.22**	**0.79**
MOS	18.27	0.59	12.74	0.61
MSA	18.44	**0.66**	11.68	0.49
MUS	19.10	0.61	5.29	0.45
RAB	15.96	0.61	5.73	0.52
RIY	16.57	0.59	13.19	0.53
SAL	15.87	0.58	7.53	0.45
SAN	18.67	**0.63**	12.56	0.73
SFX	12.92	0.55	5.69	0.52
TRI	**21.04**	0.63	8.79	0.56
TUN	13.67	0.56	4.50	0.54

Table 10: Supervised fine-tuning (SFT) per-dialect evaluation results for Qwen-2.5-7B across all MADAR-26 dialects. Bolded values indicate the top three scores within each metric column.

I 1.5B Reinforcement Learning Models' Per-Dialect Evaluation Results

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	27.63	0.72	10.70	0.71
ALE	28.20	0.71	11.08	0.72
ALG	22.80	0.71	8.92	0.69
ALX	32.16	0.75	7.77	0.72
AMM	31.65	0.75	10.65	0.73
ASW	26.16	0.73	9.03	0.72
BAG	26.72	0.72	8.60	0.71
BAS	25.33	0.71	12.72	0.71
BEI	20.84	0.65	5.50	0.68
BEN	32.07	0.73	13.10	0.72
CAI	26.33	0.72	8.00	0.72
DAM	27.26	0.73	9.78	0.72
DOH	27.90	0.73	14.26	0.73
FES	28.79	0.72	9.63	0.70
JED	24.78	0.72	12.38	0.73
JER	28.40	0.72	13.11	**0.74**
KHA	35.07	0.76	12.92	0.73
MOS	26.66	0.70	9.28	0.70
MSA	**39.71**	**0.83**	**14.57**	**0.77**
MUS	**36.43**	**0.78**	12.57	0.72
RAB	18.54	0.65	5.94	0.66
RIY	**36.76**	**0.80**	**19.27**	**0.76**
SAL	29.35	0.74	**14.88**	**0.74**
SAN	29.01	0.73	10.05	0.71
SFX	13.12	0.60	4.05	0.62
TRI	25.65	0.70	8.07	0.69
TUN	14.88	0.61	4.56	0.62

Table 11: Per-dialect evaluation results for the RL-trained Qwen-2.5-1.5B model. Bolded values indicate the top three scores (including ties) within each metric column.

J 3B Reinforcement Learning Models' Per-Dialect Evaluation Results

Dialect	Dia→En		En→Dia	
	BLEU	COMET	BLEU	COMET
Overall	34.88	0.79	11.31	0.73
ALE	37.21	0.79	**15.07**	0.74
ALG	28.10	0.77	9.61	0.69
ALX	40.88	0.82	13.69	**0.77**
AMM	38.31	0.82	14.90	0.76
ASW	33.52	0.80	8.12	0.75
BAG	33.76	0.78	10.93	0.73
BAS	32.05	0.78	12.70	0.73
BEI	29.85	0.74	6.73	0.70
BEN	39.71	0.80	**17.06**	0.73
CAI	32.73	0.80	9.27	0.73
DAM	35.26	0.80	12.97	0.74
DOH	35.43	0.80	9.90	0.75
FES	36.03	0.79	11.50	0.69
JED	31.65	0.79	8.30	0.74
JER	36.16	0.80	15.42	0.76
KHA	**42.51**	**0.83**	**16.75**	0.75
MOS	32.83	0.78	7.80	0.70
MSA	**45.06**	**0.87**	13.42	**0.80**
MUS	41.24	**0.83**	8.69	**0.77**
RAB	26.96	0.74	6.95	0.66
RIY	**44.07**	**0.86**	13.69	**0.79**
SAL	36.86	0.80	10.76	0.76
SAN	35.96	0.80	12.69	0.75
SFX	20.77	0.69	5.16	0.63
TRI	34.36	0.78	10.31	0.71
TUN	23.00	0.71	6.20	0.64

Table 12: Per-dialect evaluation results for the RL-trained Qwen-2.5-3B model. Bolded values indicate the top three scores (including ties) within each metric column.

K English→Dialect Reinforcement Learning Results Without Initial Supervised Fine-Tuning

	Qwen-2.5-1.5B		Qwen-2.5-3B	
Dialect	BLEU	COMET	BLEU	COMET
Overall	1.46	0.53	2.01	0.58
ALE	1.08	0.53	1.60	0.57
ALG	0.75	0.52	1.80	0.57
ALX	1.29	0.52	1.65	0.58
AMM	0.90	0.53	2.70	0.58
ASW	1.26	0.52	2.21	0.57
BAG	1.49	0.53	2.16	0.59
BAS	1.49	0.53	1.75	0.59
BEI	1.03	0.52	1.12	0.56
BEN	**2.17**	0.53	1.56	0.58
CAI	1.54	0.52	1.63	0.57
DAM	1.39	0.53	1.39	0.58
DOH	1.60	0.53	2.05	0.59
FES	0.99	0.52	2.54	0.57
JED	1.32	0.53	1.63	0.59
JER	1.51	0.53	1.83	0.58
KHA	**2.10**	0.54	2.53	0.59
MOS	0.96	0.52	1.51	0.57
MSA	1.44	0.53	**2.90**	0.60
MUS	1.94	0.54	**3.72**	0.61
RAB	0.95	0.51	1.08	0.54
RIY	**2.65**	0.54	**4.14**	0.60
SAL	1.96	0.53	1.88	0.58
SAN	1.54	0.53	1.38	0.58
SFX	1.15	0.51	1.06	0.54
TRI	1.65	0.53	1.25	0.57
TUN	1.09	0.51	1.15	0.55

Table 13: Evaluation results for Qwen-2.5 1.5B and 3B models trained with reinforcement learning directly from the pretrained checkpoints, without supervised fine-tuning. Bolded values indicate the top BLEU scores within each column; COMET scores are not highlighted due to heavy score concentration and frequent ties across nearly all dialects.