

FARABI at WMT 2026: Post-Hoc Channel Controllers for Low-Resource Arabic–Asian Machine Translation

Farizeh Aldabbas^{*,1,2,3} Zyad Altahan^{*,1,2} Hossam Elsafty^{1,2,3} Rafet Sifa^{1,2,3}

^{*}Equal contribution

¹Fraunhofer IAIS, Germany ²University of Bonn, Germany ³Lamarr Institute, Germany

Abstract

We describe our submission to the WMT26 Low-Resource Arabic-Asian Language Translation shared task covering ten translation directions between Arabic and five Asian languages. Our central question is whether a post-hoc controller can improve translation quality beyond a fine-tuned backbone without modifying any model weight. For the five into-Arabic directions, our primary system jointly fine-tunes facebook/nllb-200-3.3B on all source languages and augments it at inference time with per-direction NTK-Mirror controllers (fewer than 6K parameters per direction) that rescale frozen transformer output channels. For the five from-Arabic directions, the backbone is QWEN2.5-7B-INSTRUCT adapted via QLoRA, likewise augmented with NTK-Mirror. In the official evaluation, our system ranked **first** on Bengali- and Indonesian-to-Arabic, second on Hindi- and Urdu-to-Arabic, and third on English-to-Arabic, placing in the top three on all five into-Arabic directions. Post-submission analysis indicates that the weaker from-Arabic results stem primarily from a model-selection decision at submission time.

1 Introduction

Parameter-efficient fine-tuning (PEFT) has become the dominant paradigm for adapting large pre-trained models to low-resource machine translation. Methods such as LoRA (Hu et al., 2022) and QLoRA (Detrmers et al., 2023) inject trainable low-rank matrices into a frozen backbone, yielding strong MT performance at a fraction of full fine-tuning cost. Yet LoRA still introduces tens of millions of trainable parameters per adapter, far more than the task data can reliably support in the smallest low-resource settings, where each direction provides only ≈ 20 k sentence pairs.

We investigate a complementary question: can a post-hoc controller improve MT quality beyond a fine-tuned backbone, with no further gradient

flow through backbone weights? The controller we study, NTK-Mirror,¹ learns one scalar for each of a few thousand selected (layer, channel) pairs and multiplies the corresponding frozen-layer output channel by its exponential. It is fitted only after the backbone’s weights are set – fine-tuned and frozen for NLLB, or pretrained and frozen prior to the LoRA update for Qwen – so adapting a new direction never touches a backbone weight. Our answer is affirmative: on our strongest backbone NTK-Mirror delivers consistent BLEU gains for a parameter budget roughly $8,000\times$ smaller than that of a LoRA adapter.

Our system, *FARABI*, exploits three complementary strategies: (1) joint multilingual fine-tuning of NLLB-200 on all five into-Arabic directions, and QLoRA fine-tuning of QWEN2.5-7B-INSTRUCT for from-Arabic; (2) NTK-Mirror controllers trained post-hoc on frozen backbone representations to correct systematic representation biases with negligible parameter overhead; and (3) a retrieval-augmented second-pass adapter, submitted as a contrastive system for from-Arabic.

Our work makes three contributions:

- We fine-tune two backbone models (NLLB-200, Qwen2.5-7B) under different low-resource adaptation settings to produce MT systems for ten Arabic–Asian directions.
- We evaluate NTK-Mirror as a parameter-minimal post-hoc controller for low-resource MT, adding only ≈ 5 K trainable gate scalars over frozen backbone representations.
- We report both successful into-Arabic results and a negative from-Arabic model-selection finding, showing that direction-specific architecture choice can dominate the benefit of lightweight adaptation.

The paper is organised around three questions, examined separately for each half of the task:

¹<https://github.com/leochlon/ntkmirror>

(1) does a post-hoc channel controller improve on an already fine-tuned backbone? (2) does a single multitask model generalise better than per-direction tuning? and (3) how much of our from-Arabic gap to the leading systems is attributable to backbone choice rather than to the adaptation method? Sections 6.2 and 6.3 take up the first two questions for into- and from-Arabic respectively, and Section 6.4 answers the third with post-submission experiments.

2 Related Work

Large multilingual LLMs have been applied to low-resource MT via zero-shot prompting (Vilar et al., 2023; Zhu et al., 2024) and lightweight adaptation (Jiao et al., 2023; Xu et al., 2024). QLoRA (Dettmers et al., 2023) has become the standard for adapting large models to specialised tasks with limited data. NTK-Mirror is complementary to both: it leaves every backbone weight untouched and instead learns a small set of per-channel scales on the outputs of frozen layers, requiring no backward pass through those weights.

A conceptually related channel-level intervention is IA³ (Liu et al., 2022), which scales activations via learned vectors ($\tilde{\mathbf{h}}^{(l)} = \mathbf{h}^{(l)} \odot \mathbf{1}$) injected during fine-tuning as a primary adaptation mechanism. We instead ask whether similar gating scalars can be optimized *post-hoc* over frozen backbone representations, independently of the primary adaptation procedure; applying such parameter-minimal, post-hoc controllers to correct residual biases in MT is, to our knowledge, novel.

Retrieval-augmented generation for MT dates to Gu et al. (2018) and has been revisited for LLMs by Agrawal et al. (2023), whose in-context exemplar selection our contrastive system follows.

3 Methodology

3.1 Task Description

The WMT26 Low-Resource Arabic-Asian MT shared task (Laskar et al., 2026) covers ten directions between Arabic and five Asian languages (English, Hindi, Bengali, Indonesian, Urdu): five *into* Arabic and five *from* Arabic. The official task codes are 1A–1E for EN/HI/BN/ID/UR→AR and 2A–2E for AR→EN/HI/BN/ID/UR. For readability we name directions explicitly throughout the text and retain the codes only in table row labels. Each direction provides a 1,500-sentence challenge

test set, evaluated primarily with BLEU (Papineni et al., 2002) and chrF (Popović, 2015).

3.2 NTK-Mirror Controllers

Motivation. A LoRA-fine-tuned backbone learns task-specific transformations in the weight space of individual projection matrices. However, at inference time certain channels in the residual stream may remain systematically over- or under-active for a specific translation direction, a residual bias that low-rank updates cannot fully correct without overfitting on small data. NTK-Mirror targets this directly by learning per-channel scales in the output space of frozen layers. No backbone weight is modified: the intervention is a multiplicative mask applied at activation time, leaving the model bit-for-bit identical between directions.

Method. Let $\mathbf{h}_{l,t} \in \mathbb{R}^D$ be the output hidden state at layer l and token position t . NTK-Mirror introduces a log-gate scalar $s_{l,c} \in \mathbb{R}$ for each selected (layer, channel) pair and applies:

$$\mathbf{h}'_{l,t,c} = e^{s_{l,c}} \cdot \mathbf{h}_{l,t,c} \quad (1)$$

The exponential parameterisation preserves sign structure and keeps the optimisation landscape smooth; $s_{l,c}=0$ (identity) is the natural initialisation. Figure 1 illustrates how the two adaptation mechanisms interact within a single transformer layer.

Gate selection. Training all $L \times D$ gates ($28 \times 4096 \approx 115\text{k}$ for Qwen2.5-7B) introduces a risk of overfitting on 20k sentence pairs. We therefore restrict to the top- G most informative gates. Gate importance is scored by mean activation magnitude on a held-out subset, avoiding the added cost and numerical instability of a full backward pass through the frozen backbone. The top $G=5\,000$ gates by this criterion are retained; all others remain fixed at zero.

Parameter efficiency. Table 1 places NTK-Mirror in context. 5 000 scalars is roughly $8,000 \times$ fewer parameters than the QLoRA adapter, yet the method requires no additional backward pass through backbone layers.

Joint application with LoRA (Qwen only). For the Qwen backbone, which is adapted via LoRA rather than by updating its own weights, the NTK-Mirror controller is trained on the *raw, pre-LoRA frozen backbone*, not on the LoRA-merged weights,

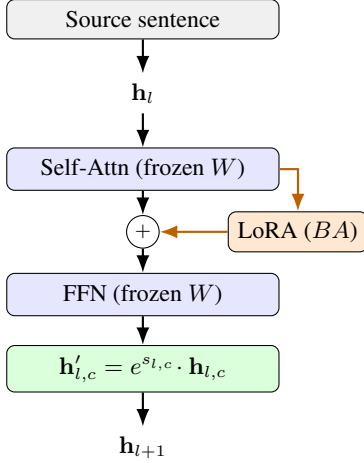


Figure 1: One transformer layer showing the Qwen+QLoRA+NTK-Mirror system. Frozen weights W (blue, 4-bit NF4 quantised) are unchanged throughout. QLoRA adapters A , B (orange, rank 16, the low-rank BA update shown in the box) are trained in stage 1. NTK-Mirror gates $s_{l,c}$ (green) are trained post-hoc in stage 2 on the raw frozen backbone and applied jointly at inference. For the NLLB into-Arabic system the LoRA branch is absent; NTK-Mirror gates are applied directly to the fine-tuned frozen backbone.

Method	Trainable	% of model
Full fine-tuning	7 B	100%
QLoRA (ours)	≈ 40 M	0.9%
NTK-Mirror (ours)	≈ 5 K	$< 0.001\%$
QLoRA + NTK-Mirror	≈ 40 M + 5 K	0.9%

Table 1: Trainable parameter counts per adaptation method. NTK-Mirror is trained separately on the frozen backbone, either alone (NLLB) or alongside a LoRA adapter applied at inference (Qwen).

so that gate scores reflect the backbone’s intrinsic representation geometry; at inference, the LoRA adapter and NTK-Mirror gates are applied together. NLLB has no LoRA adapter, so its controller is instead trained directly on the already fine-tuned frozen backbone (Section 4.1).

4 System Description

The shared task accepts one PRIMARY submission per direction together with two contrastive submissions, C1 and C2; only the primary is ranked. Table 2 records which system we placed in each slot for each direction, and the subsections below describe those systems in turn.

FARABI uses two backbones. **Into-Arabic** primary and C1 systems use facebook/nllb-200-3.3B (Costa-jussà et al., 2022), a seq2seq model with strong multilingual

coverage; **from-Arabic** uses QWEN2.5-7B-INSTRUCT (Hui et al., 2024), an instruction-tuned LLM with broad Arabic/Asian script coverage. In both cases, per-direction NTK-Mirror controllers are trained post-hoc and applied at inference alongside the direction-specific adaptation – on the fine-tuned backbone directly for NLLB, or on the raw pre-LoRA backbone jointly with the LoRA adapter for Qwen (Section 3.2).

Slot	Description
PRIMARY	All five into-Arabic directions: NLLB-200 jointly fine-tuned on all five + per-direction NTK-Mirror controller. $AR \rightarrow EN$, $AR \rightarrow ID$: Qwen2.5-7B-Instruct QLoRA + NTK-Mirror. $AR \rightarrow HI$, $AR \rightarrow BN$, $AR \rightarrow UR$: NLLB-200 zero-shot, selected by COMET-QE on the challenge test set.
C1	Into-Arabic: NLLB-200 fine-tuned (no NTK-Mirror). From-Arabic: Qwen2.5-7B-Instruct QLoRA only (no NTK-Mirror).
C2	Into-Arabic: Qwen2.5-7B-Instruct + QLoRA adapter. From-Arabic: Retrieval-augmented second-pass adapter with hybrid BM25+E5 retrieval (top- $k=2$), Section 4.4.

Table 2: Submitted system variants. All Qwen-based systems use Qwen2.5-7B-Instruct with 4-bit QLoRA.

4.1 Into-Arabic Primary: NLLB-200 Fine-Tuned + NTK-Mirror

We jointly fine-tune NLLB-200 on the official training pairs for all five into-Arabic directions (≈ 102 k sentences: EN, HI, BN, ID, UR $\rightarrow$ AR). Joint training allows the model to share multilingual representations while maintaining a single deployment checkpoint; we observed consistent improvements across all directions compared to single-pair models (see Section 6.2). Training details are in Section 5.

After fine-tuning, per-direction NTK-Mirror controllers ($G=5\,000$ gates) are trained on the frozen fine-tuned model and applied at inference, producing the submitted FARABI-PRIMARY for all five into-Arabic directions. The NLLB fine-tuned model without controllers is submitted as FARABI-C1 for into-Arabic. FARABI-C2 for into-Arabic is the Qwen2.5-7B+QLoRA system.

4.2 From-Arabic C1: QLoRA Baseline

The backbone is loaded in 4-bit NF4 quantisation with double quantisation and bfloat16 compute dtype. LoRA adapters ($r=16$, $\alpha=32$, dropout 0.05)

are applied to all seven projection matrices in every transformer block, yielding ≈ 40 M trainable parameters atop the frozen backbone. Training details are deferred to Section 5. A direction-specific instruction prompt is prepended to each example:

```
You are a machine translation system.
Translate faithfully from {src} to {tgt}.
Return only the translation in {tgt},
with no explanation. Output ONLY {script}
script.
Source ({src}): {text}
Translation ({tgt}):
```

We also train a **multitask** adapter on all ten directions jointly for comparison (Section 6.3).

4.3 From-Arabic Primary: QLoRA + NTK-Mirror

For from-Arabic, our primary submission extends the C1 QLoRA adapter with the NTK-Mirror controller. A per-direction controller ($G=5\,000$ gates) is trained on the *raw frozen Qwen backbone* (without the LoRA adapter merged) and applied jointly with LoRA at inference. This system is our FARABI-PRIMARY for AR $\rightarrow$ EN and AR $\rightarrow$ ID. For AR $\rightarrow$ HI, AR $\rightarrow$ BN and AR $\rightarrow$ UR, a reference-free COMET-QE (quality estimation, no reference translation needed) comparison on the challenge test set showed an NLLB-200 zero-shot baseline outperforming the Qwen system, so those predictions were adopted as primary instead — a decision we revisit in Section 6.4.

4.4 C2: Retrieval-Augmented Second Pass

FARABI-C2 extends the C1 adapter with a two-stage retrieval-augmented SFT pipeline drawn from Agrawal et al. (2023).

Hybrid retrieval. We build a retrieval index over the training corpus combining BM25 (Robertson et al., 1994) (weight 0.4) and a dense encoder, *intfloat/multilingual-e5-base* (Wang et al., 2022) (weight 0.6). The final score is restricted to the same language pair to avoid cross-lingual noise.

Second-pass SFT. Starting from the C1 checkpoint, each training example is augmented with the top- $k=2$ retrieved parallel examples injected as in-context exemplars. A 25% retrieval-dropout rate during training improves robustness to retrieval failure at test time.

AbstentionTrainer. The second-pass loss is computed by a custom *AbstentionTrainer* intended to

discourage fluent-but-unfaithful output. It down-weights (factor 0.1) any in-batch sample whose per-token cross-entropy exceeds the 80th percentile, and amplifies the loss ($\times 2.0$) on synthetic negatives built by corrupting reference translations with hallucinated numbers, wrong names or omitted content. At inference, per-token mean log-probability acts as a confidence proxy: candidates below -1.5 trigger a retrieval-retry, and a post-generation checker verifies numeric tokens and hashtags. We describe this component because it was part of the submitted C2 system, but we did not ablate it separately and therefore make no claim about its individual contribution; the C2 numbers reported in Section 6.3 reflect retrieval and abstention jointly.

5 Experimental Setup

5.1 Data

The official dataset (Laskar et al., 2026) provides approximately 20 000–21 000 training sentence pairs per language pair, with development and devtest sets of $\approx 1\,000$ and 500 sentences respectively (Table 3). No external data was used beyond the official corpus.

Language Pair	Train	Dev / Devtest
Ar-En	21,000	1,000 / 500
Ar-Hi	20,100	$\approx 1,000$ / 500
Ar-Bn	20,500	$\approx 1,000$ / 500
Ar-Id	20,400	$\approx 1,000$ / 500
Ar-Ur	20,300	$\approx 1,000$ / 500

Table 3: Training corpus sizes per language pair (both directions use the same aligned corpus).

5.2 Evaluation

We evaluate on the official devtest set (500 sentences per direction) using *sacrebleu* (Post, 2018) for BLEU (tokenisation: 13a for non-Arabic targets, intl otherwise) and *unbabel-comet* (*Unbabel/wmt22-comet-da*) for neural evaluation. Reference-free COMET-QE (Rei et al., 2022) (*Unbabel/wmt22-cometkiwi-da*) was used for final system selection on the 1,500-sentence challenge test set, where no references were available. All experiments ran on two NVIDIA RTX 6000 Ada GPUs (48 GB each).

5.3 Training Details

NLLB-200. Training uses $\text{lr}=2 \times 10^{-5}$, effective batch size 32 (physical batch $2 \times$ gradient accumulation 16), *bfloat16* precision, 5 epochs, and

BLEU-based checkpoint selection. Post-hoc NTK-Mirror controllers are optimised with AdamW for 240 steps at $\text{lr}=5\times 10^{-3}$.

Qwen2.5-7B. QLoRA training uses AdamW (fused), a cosine learning-rate schedule, warmup ratio 0.03, peak $\text{lr}=1\times 10^{-4}$, weight decay 0.01, gradient clipping at 1.0, and gradient checkpointing. Effective batch size is 32. We train for 3 epochs with sequence length capped at 1 536 tokens. For the retrieval C2 system, we continue training for 1 epoch at $\text{lr}=10^{-4}$ (max sequence length 1 792 tokens).

6 Results

Full devtest scores appear in Appendix Tables 6 and 7.

6.1 Official Primary Results

Table 4 reports the official challenge-test leaderboard results for the FARABI primary submission. FARABI ranked in the top three for all five into-Arabic tasks, including first place on BN→AR and ID→AR. Conversely, it was substantially weaker on most from-Arabic tasks. The two halves therefore warrant separate treatment, which the next two subsections give them; Section 6.4 then shows that the from-Arabic gap largely reflects the submitted backbone and fine-tuning choices.

6.2 Into-Arabic Systems

Joint vs. per-direction fine-tuning. Fine-tuning NLLB-200 jointly on all five into-Arabic directions improved consistently over single-direction models on every metric, which is why the joint checkpoint serves as the baseline for everything that follows.

NTK-Mirror on the joint backbone (submitted primary). NTK-Mirror provides modest but consistent gains on this backbone. BLEU improves by +0.2 to +0.6 across all five directions (e.g., 20.53 to 21.04 on EN→AR), chrF2 improves consistently across all directions (+0.5 to +0.7), and COMET shows modest gains (+0.3 to +0.7). A paired-bootstrap significance analysis (sacrebleu, 1 000 resamples) on an independently retained run that reproduces the C1 baseline scores found significant NTK-Mirror improvements for four of five directions (EN→AR $p=0.008$, BN→AR $p=0.036$, ID→AR $p=0.006$, UR→AR $p=0.008$); HI→AR was positive but not significant ($p=0.082$). In an independent gate-count sensitivity run

($G \in \{500, 1\,000, 2\,500, 5\,000, 10\,000, 20\,000\}$ on EN→AR, Appendix Table 8) shows little sensitivity to G within that pass: the $G=5\,000$ setting (20.62 BLEU) is effectively tied with that pass’s own no-controller baseline (20.64), and BLEU stays within this narrow band across the whole range. We therefore treat $G=5\,000$ as a reasonable fixed operating point rather than a tuned optimum.

A Qwen backbone for comparison (C2). As a contrastive check we applied the same controller to a Qwen+QLoRA backbone on the same five directions. Unlike the NLLB primary, the effect here is small and direction-dependent rather than consistently positive. Earlier devtest runs for EN→AR and UR→AR inadvertently used the from-Arabic test set (a direction mismatch in the scoring script); corrected into-Arabic evaluations are reported in Table 6. After correction, NTK-Mirror is near-neutral on EN→AR (−0.40 BLEU, +0.11 chrF2, +0.32 COMET) and shows small consistent gains on UR→AR (+0.45 BLEU, +0.43 chrF2, +0.42 COMET). For HI→AR the gain is small (+0.23 BLEU), while for BN→AR the controller slightly decreases BLEU (−0.47). Across all five directions the NLLB-FT primary outperforms the Qwen+NTK ablation on COMET (e.g., 87.88 vs. 86.92 on EN→AR), suggesting the NLLB backbone produces more semantically faithful output even where the two are comparable on BLEU.

6.3 From-Arabic Systems

Multitask vs. per-direction adapters. A single QLoRA adapter trained jointly on all ten directions beats per-direction adapters on all five from-Arabic directions (+2.27 to +5.56 BLEU), suggesting that joint multilingual training is beneficial in this setting. At submission time we had trained NTK-Mirror controllers only on the per-direction adapters, so this configuration was not submitted as primary and was evaluated on devtest only.

NTK-Mirror on the Qwen adapter (submitted primary for AR→EN and AR→ID). NTK-Mirror adds +0.08 BLEU on AR→EN and +3.85 BLEU on AR→ID, consistent with the utility of per-direction channel scaling when the underlying adapter is itself direction-specific.

Prompt template. The prompt of Section 4.2 was not tuned per direction; it is identical across all five from-Arabic directions apart from the language names and script warning. A check on AR→EN

Task	Direction	Rank	Teams	BLEU	chrF2	TER↓	COMET*	EmbSim [†]	ΔBLEU↓	Best team
1A	EN→AR	3	9	20.57	53.34	69.45	84.04	85.03	1.88	NLP-IIT Patna
1B	HI→AR	2	8	18.62	51.53	72.01	83.44	84.30	1.24	NLP-IIT Patna
1C	BN→AR	1	6	17.93	50.73	72.71	83.04	84.03	0.00	FARABI
1D	ID→AR	1	4	18.72	51.24	71.70	83.39	84.23	0.00	FARABI
1E	UR→AR	2	6	18.32	50.94	72.29	83.26	84.18	0.68	NLP-IIT Patna
2A	AR→EN	9	10	22.76	54.79	70.31	80.47	79.15	11.08	OpenBracket
2B	AR→HI	10	10	21.84	47.20	68.20	68.73	83.55	7.44	NLP-IIT Patna
2C	AR→BN	7	7	10.86	44.76	80.07	77.12	81.16	6.24	NERDS-NITS
2D	AR→ID	3	4	23.95	56.65	67.75	83.75	84.50	6.57	NERDS-NITS
2E	AR→UR	5	6	18.31	44.63	71.81	74.01	81.37	7.59	SLPG_FJWU

Table 4: Official primary-submission results for FARABI on the WMT26 challenge test set (Laskar et al., 2026). ΔBLEU = gap to the best primary system. *COMET reported on 0–100 scale by the official evaluator. [†]EmbSim = mBERT cosine similarity. Bold rank = first place.

and AR→ID (Appendix Table 9) shows that removing its output-format constraints costs 15–20 BLEU, so the constraints are necessary rather than cosmetic.

Retrieval augmentation (C2). Retrieval augmentation (**oracle**²) helps AR→BN and AR→UR by up to 6 BLEU but hurts AR→ID (−7.3 BLEU). We hypothesize, without direct verification, that the AR→ID training corpus has lower sentence-level similarity to test inputs than the other four directions, so the 25% retrieval-dropout rate did not prevent misleading retrieved context; we leave a per-direction retrieval-quality analysis to future work.

6.4 Diagnosing the From-Arabic Gap

The from-Arabic results in Table 4 admit two explanations: either post-hoc channel gating fails out of Arabic, or the backbones we submitted were simply the wrong ones. To separate the two, we repeated the into-Arabic recipe on the other half of the task after the deadline, jointly fine-tuning NLLB-200 on all five from-Arabic directions (≈102k sentence pairs) and training per-direction NTK-Mirror controllers on top.

Table 5 supports the second explanation. Joint from-Arabic fine-tuning improves over the submitted primary on all five directions: +3.7 BLEU on AR→EN, +5.2 on AR→HI, +4.9 on AR→BN, +2.9 on AR→ID and +6.6 on AR→UR. COMET improves by 3.2–11.2 points, with the largest gain on AR→UR (74.01 → 80.18) and on AR→HI, where the zero-shot primary was worst (68.73 → 72.97). On AR→ID the fine-tuned model exceeds

even the best submitted system on COMET (85.67 vs. 84.11). Adding NTK-Mirror on top changes little in either direction (<0.5 BLEU everywhere), suggesting that its effect depends on the backbone and adaptation regime rather than being uniform across configurations. The from-Arabic weakness in the official results therefore appears to stem primarily from model selection. Fine-tuning on ~100k sentence pairs of the official data substantially narrows the gap to the top systems independently of the adaptation method, though we note that these devtest figures are compared against challenge-test figures for the submitted and best systems, so the margins are indicative rather than exact.

7 Discussion

Parameter-minimal adaptation is viable.

Fewer than 6K log-gate scalars trained post-hoc on a frozen backbone provide a parameter-minimal refinement with negligible additional trainable parameters. On the jointly fine-tuned NLLB into-Arabic backbone, NTK-Mirror yields consistent BLEU gains across all five directions, with improvements statistically significant in four of five directions in our paired-bootstrap analysis. Results on other backbone and adaptation configurations are more direction-dependent, indicating that the controller’s benefit depends on the underlying adaptation regime.

Architecture alignment matters as much as adaptation method.

The official leaderboard shows a sharp asymmetry: into-Arabic (NLLB-200+NTK-Mirror) took two first-place, two second-place and one third-place result, while from-Arabic was substantially weaker for a concrete reason —

²Oracle: retrieval is drawn from a test-time reference-parallel index and is not a deployable configuration.

System	BLEU	chrF2	COMET*
<i>AR→EN</i>			
Submitted (Qwen+NTK)	22.76	54.79	80.47
Best submitted system	33.83	59.72	82.49
NLLB-FT (post-sub.)	26.47	54.19	83.69
NLLB-FT+NTK (post-sub.)	26.51	54.45	83.73
<i>AR→HI</i>			
Submitted (NLLB zero-shot)	21.84	47.20	68.73
Best submitted system	29.28	54.73	72.34
NLLB-FT (post-sub.)	27.05	52.60	72.97
NLLB-FT+NTK (post-sub.)	26.20	51.85	72.84
<i>AR→BN</i>			
Submitted (NLLB zero-shot)	10.86	44.76	77.12
Best submitted system	17.10	51.83	78.69
NLLB-FT (post-sub.)	15.72	51.06	80.85
NLLB-FT+NTK (post-sub.)	15.88	50.49	80.48
<i>AR→ID</i>			
Submitted (Qwen+NTK)	23.95	56.65	83.75
Best submitted system	30.51	59.82	84.11
NLLB-FT (post-sub.)	26.84	57.67	85.67
NLLB-FT+NTK (post-sub.)	26.28	57.18	85.39
<i>AR→UR</i>			
Submitted (NLLB zero-shot)	18.31	44.63	74.01
Best submitted system	25.89	51.06	76.26
NLLB-FT (post-sub.)	24.88	50.37	80.18
NLLB-FT+NTK (post-sub.)	24.95	50.34	80.06

Table 5: Post-submission from-Arabic analysis. Submitted and best-system rows are official challenge-test scores (Laskar et al., 2026); NLLB-FT rows are 500-sentence devtest scores. Bold = best post-submission devtest score per direction. *COMET on 0–100 scale.

for AR→HI, AR→BN and AR→UR the submitted primary was a *zero-shot* NLLB-200 model with no from-Arabic supervision at all. Section 6.4 quantifies what that choice cost: the post-submission jointly fine-tuned NLLB system improves over the submitted primary by +2.9 to +6.6 BLEU and 3.2–11.2 COMET across all five directions. Taken together, the two halves suggest that picking a backbone already aligned with the target direction mattered more than any of the lightweight components we layered on top of it.

8 Conclusion

We presented FARABI, a multi-system submission to the WMT26 Low-Resource Arabic-Asian Machine Translation shared task. Our primary contribution is an empirical study of NTK-Mirror controllers as a post-hoc, parameter-minimal complement to standard MT fine-tuning. Our jointly fine-tuned NLLB-200 system, augmented with per-direction NTK-Mirror controllers, achieved two first-place, two second-place, and one third-place results across the five into-Arabic directions; with fewer than 6K trainable scalars, the controller itself delivers consistent but modest BLEU gains

on top of that backbone (+0.2–0.6 across all five directions). For from-Arabic, NTK-Mirror complements Qwen2.5-7B-Instruct with QLoRA, with gains that are smaller and direction-dependent. Across both halves of the task, backbone selection and fine-tuning – not the post-hoc controller – remain the dominant factor in translation quality; NTK-Mirror is best understood as a parameter-minimal refinement whose benefit depends on the underlying backbone and adaptation regime, rather than a substitute for strong backbone adaptation.

Acknowledgements

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence. We thank the WMT26 organisers and the Elyadata team for providing the Arabic-Asian MT datasets.

References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.
- Marta R. Costa-jussà et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.
- Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParrotT: Translating during chat using large language models tutored with human feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020.

- Sahinur Rahman Laskar, Firoj Alam, Bishwaraj Paul, Irfan Ahmad, Maya Silvi Lydia, and Pankaj Kundan Dadure. 2026. Findings of the WMT 2026 shared task on low-resource arabic-asian language translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohhta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- 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.
- Ricardo Rei, Ana C. Farinha, José G. C. de Souza, Pedro Henrique Ramos, André F. T. Martins, Luísa Coheur, and Alon Lavie. 2022. Cometkiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *International Conference on Learning Representations*.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.

A Full Devtest Ablation Results

Table 6 reports devtest scores for all system variants described in the paper, including C2 (retrieval-augmented), multitask, and NTK-Mirror configurations.

Sub-task	System	BLEU	chrF	TER	METEOR	COMET	Submission
<i>Into Arabic: Qwen2.5-7B+QLoRA C2 ablation (submitted primary/C1 were NLLB-200; see Table 7)</i>							
1A EN→AR	Qwen+QLoRA	16.93	50.95	74.52	36.57	86.60	C2
	Qwen+QLoRA+Retrieval	15.65	48.69	77.54	33.54	85.20	–
	Qwen+NTK	16.53	51.06	74.39	—	86.92	–
1B HI→AR	Qwen+QLoRA	12.46	46.19	82.35	30.19	83.90	C2
	Qwen+QLoRA+Retrieval	12.59	44.77	83.79	28.66	83.10	–
	Qwen+NTK	12.69	44.64	84.22	28.56	83.00	–
1C BN→AR	Qwen+QLoRA	12.59	45.97	81.60	29.93	84.20	C2
	Qwen+QLoRA+Retrieval	11.86	44.15	85.39	27.54	83.20	–
	Qwen+NTK	12.12	44.07	86.09	28.41	82.80	–
1D ID→AR	Qwen+QLoRA	11.69	45.01	85.46	27.68	84.40	C2
	Qwen+QLoRA+Retrieval	10.12	39.53	88.33	24.48	78.40	–
	Qwen+NTK	12.13	44.66	85.57	28.74	83.50	–
1E UR→AR	Qwen+QLoRA	11.10	44.99	87.64	28.04	84.70	C2
	Qwen+QLoRA+Retrieval	11.85	44.28	85.47	27.83	83.90	–
	Qwen+NTK	11.55	45.42	84.53	—	85.12	–
<i>From Arabic</i>							
2A AR→EN	Multitask QLoRA	26.57	56.81	65.05	49.65	83.40	–
	QLoRA (C1)	24.30	56.04	69.07	47.31	83.10	C1
	QLoRA + Retrieval (C2)	23.00	50.38	72.39	43.17	78.40	C2
	QLoRA + NTK-Mirror (Primary)	24.38	55.84	67.32	47.09	83.30	Primary
2B AR→HI	Multitask QLoRA	19.56	45.53	76.66	39.75	65.20	–
	QLoRA (C1)	14.79	38.86	78.44	33.78	61.70	C1
	QLoRA + Retrieval (C2)	18.31	41.71	77.42	35.88	61.80	C2
	QLoRA + NTK-Mirror	18.96	45.66	76.89	39.42	66.40	–
	NLLB-200 zero-shot [†]	–	–	–	–	–	Primary
2C AR→BN	Multitask QLoRA	14.17	40.64	81.41	30.04	72.60	–
	QLoRA (C1)	8.61	37.52	87.41	22.47	68.10	C1
	QLoRA + Retrieval (C2, oracle)	14.68	47.90	80.31	33.47	76.50	C2
	QLoRA + NTK-Mirror	11.07	44.73	85.17	28.27	74.70	–
	NLLB-200 zero-shot [†]	–	–	–	–	–	Primary
2D AR→ID	Multitask QLoRA	21.09	46.98	73.94	43.68	76.30	–
	QLoRA (C1)	17.69	43.88	76.22	39.31	74.60	C1
	QLoRA + Retrieval (C2)	10.40	30.78	83.95	27.27	61.80	C2
	QLoRA + NTK-Mirror (Primary)	21.54	47.19	72.75	43.88	77.00	Primary
2E AR→UR	Multitask QLoRA	21.16	47.32	73.52	43.72	76.30	–
	QLoRA (C1)	17.95	44.06	75.42	39.46	74.90	C1
	QLoRA + Retrieval (C2, oracle)	20.94	46.93	72.26	42.88	76.00	C2
	QLoRA + NTK-Mirror	20.76	46.91	72.82	42.92	77.00	–
	NLLB-200 zero-shot [†]	–	–	–	–	–	Primary

Table 6: Full devtest ablation results (500 sentences). Bold = best score per sub-task. COMET on 0–100 scale.

[†]NLLB-200 zero-shot was the actual primary submission for AR→HI, AR→BN and AR→UR; devtest scores were not computed pre-submission. The 1A and 1E Qwen+NTK rows previously showed from-Arabic evaluation scores due to a test-set mismatch in the scoring script; corrected into-Arabic scores are now reported. Rows marked *oracle* use test-time reference retrieval and are not deployable configurations.

B Devtest Results: All System Variants

Table 7 reports devtest BLEU, chrF, TER, and COMET for all submitted and post-submission systems.

Direction	System	n-gram			Neural
		BLEU	chrF	TER	COMET
Into Arabic (Task 1): NLLB-200 jointly fine-tuned (submitted systems)					
1A EN→AR	NLLB-FT (C1)	20.53	53.95	69.90	87.17
	NLLB-FT+NTK (Primary)	21.04	54.62	67.56	87.88
1B HI→AR	NLLB-FT (C1)	18.42	52.06	71.68	85.70
	NLLB-FT+NTK (Primary)	18.63	52.64	69.97	86.25
1C BN→AR	NLLB-FT (C1)	18.02	51.74	71.49	85.76
	NLLB-FT+NTK (Primary)	18.36	52.45	70.87	86.31
1D ID→AR	NLLB-FT (C1)	18.72	51.98	71.88	85.43
	NLLB-FT+NTK (Primary)	19.37	52.61	69.81	86.00
1E UR→AR	NLLB-FT (C1)	17.80	51.31	71.92	86.57
	NLLB-FT+NTK (Primary)	18.30	51.81	70.36	86.99
Into Arabic (Task 1): Qwen2.5-7B+QLoRA contrastive ablation (C2)					
1A EN→AR	Qwen+QLoRA	16.93	50.95	74.52	86.60
	Qwen+NTK	16.53	51.06	74.39	86.92
1B HI→AR	Qwen+QLoRA	12.46	46.19	82.35	83.90
	Qwen+NTK	12.69	44.64	84.22	83.00
1C BN→AR	Qwen+QLoRA	12.59	45.97	81.60	84.20
	Qwen+NTK	12.12	44.07	86.09	82.80
1D ID→AR	Qwen+QLoRA	11.69	45.01	85.46	84.40
	Qwen+NTK	12.13	44.66	85.57	83.50
1E UR→AR	Qwen+QLoRA	11.10	44.99	87.64	84.70
	Qwen+NTK	11.55	45.42	84.53	85.12
From Arabic (Task 2): Qwen2.5-7B+QLoRA submitted systems					
2A AR→EN	QLoRA (C1)	24.30	56.04	69.07	83.10
	QLoRA+NTK (Primary)	24.38	55.84	67.32	83.30
2B AR→HI	QLoRA (C1) [†]	14.79	38.86	78.44	61.70
2C AR→BN	QLoRA (C1) [†]	8.61	37.52	87.41	68.10
2D AR→ID	QLoRA (C1)	17.69	43.88	76.22	74.60
	QLoRA+NTK (Primary)	21.54	47.19	72.75	77.00
2E AR→UR	QLoRA (C1) [†]	17.95	44.06	75.42	74.90
From Arabic (Task 2): NLLB-200 fine-tuned post-submission					
2A AR→EN	NLLB-FT	26.47	54.19	–	83.69
	NLLB-FT+NTK	26.51	54.45	–	83.73
2B AR→HI	NLLB-FT	27.05	52.60	–	72.97
	NLLB-FT+NTK	26.20	51.85	–	72.84
2C AR→BN	NLLB-FT	15.72	51.06	–	80.85
	NLLB-FT+NTK	15.88	50.49	–	80.48
2D AR→ID	NLLB-FT	26.84	57.67	–	85.67
	NLLB-FT+NTK	26.28	57.18	–	85.39
2E AR→UR	NLLB-FT	24.88	50.37	–	80.18
	NLLB-FT+NTK	24.95	50.34	–	80.06

Table 7: Devtest results (500 sentences per direction). Bold = best per direction within each system block. COMET on 0–100 scale. [†]NLLB-200 zero-shot was selected as primary for AR→HI, AR→BN and AR→UR.

C Gate-Count and Prompt Ablations

G	BLEU	chrF2	COMET
500	20.47	52.27	87.42
1 000	20.54	51.97	87.41
2 500	20.66	53.30	87.38
5 000 (used)	20.62	51.54	87.34
10 000	20.57	52.23	87.37
20 000	20.77	52.34	87.20
no controller	20.64	52.72	87.47

Table 8: Gate-count (G) sensitivity, NLLB EN→AR devtest, run as a separate evaluation pass; absolute values differ slightly from the primary devtest figures in Table 7. Performance is broadly stable across $G=500$ – $20,000$; the $G=5,000$ operating point is effectively tied with this pass’s own no-controller baseline (20.62 vs. 20.64 BLEU).

Sub-task	Prompt	BLEU	COMET
2A AR→EN	Full (production)	24.40	83.08
	Minimal (no format constraints)	5.87	66.68
2D AR→ID	Full (production)	25.15	85.47
	Minimal (no format constraints)	8.76	75.63

Table 9: Prompt ablation, Qwen+QLoRA (C1) devtest. The minimal prompt drops the explanation ban, script warning, and formatting-preservation line, keeping only a bare translation request.

The minimal prompt used for Table 9, shown here in full for comparison against the production prompt in Section 4.2:

```
Translate from {src} to {tgt}: {text}
{tgt}:
```