CUNI-AR Team at WMT26 General Translation Task for Egyptian Arabic

Josef Jon[1] **Rawan Bondok**[2] **Miroslav Hrabal**[1] **Ondřej Bojar**[1]
[1]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
jon@ufal.mff.cuni.cz
[2]TU Darmstadt, Germany
rawanessam34@gmail.com

Abstract

We describe the Charles University CUNI-AR submission to the WMT26 General Translation shared task for English into Egyptian Arabic. Our pipeline is built around GEMBA (LLM-as-judge) quality estimator that scores translations independently for *accuracy* and *dialect authenticity*. We use it throughout our pipeline: to filter training data, to construct preference pairs for DPO, to select best checkpoints, and to pick the best hypothesis per segment at inference time. Our primary translation model is Gemma-4-12B-it finetuned via QLoRA, we also finetune Jais-2-8B-Chat and Aya-expanse-8B. Gemma-4-12B-it provides a high-quality Egyptian Arabic translation, so SFT yields only marginal quality gains in automatic metrics; DPO is able to improve the metrics it optimizes for, but the real gain in translation quality is to be assessed by human evaluation. The final submission selects the best-scoring translation per segment from 31 candidate checkpoints, with the unmodified base model contributing 31% of the output.

1 Introduction

Translating into Arabic dialectal varieties poses challenges that go beyond those of standard machine translation into high-resource languages. Both spoken language and written norms differ across varieties, and the boundaries between dialects are gradient rather than discrete. Standard quality-estimation models such as COMET-QE treat Arabic as a single language and have no mechanism to assess whether a translation is idiomatic in the desired variety, making it harder to automatically evaluate the quality of the translation. Dialect-differentiated test sets are scarce, so even reference-based metrics do not solve the issue.

Our submission (team **CUNI-AR**) is built around a quality estimation using LLM as a judge (GEMBA), which scores translations independently for *accuracy* (meaning preservation) and *dialect authenticity* (idiomatic correctness for the target variety). We use this signal at every stage of the pipeline: to filter training data, to construct preference pairs for Direct Preference Optimization (DPO), to select checkpoints after finetuning, and to perform segment-level best translation selection at inference time. For the translation models themselves we finetune Gemma-4-12B-it (Team, 2026), Jais-2-8B-Chat (Anwar et al., 2026) and Aya-expanse-8B (Dang et al., 2024) using QLoRA (Dettmers et al., 2023) both for supervised finetuning (SFT) and direct/contrastive preference optimization (DPO/CPO).

The main findings of our work are as follows. First, GEMBA with a Gemma-4-31B-it judge correlates strongly with WMT25 human scores (Pearson $r = 0.916$ overall), validating its use as a quality signal. Second, supervised finetuning on GEMBA-filtered synthetic data (WMT23 English sources machine translated to multiple Arabic dialects) yields only marginal or negative gains on chrF and GEMBA, indicating that Gemma-4-12B-it is already a strong Egyptian Arabic translator out of the box (at least compared to other models used to create the synthetic data). With the exception of the Alexandria dataset, all the other datasets we used for finetuning decreased both GEMBA and ChrF scores on the WMT25 test set. Finally, DPO on preference pairs that include both intra-dialect — choosing the best-scoring translation among multiple options, and cross-dialect signal — pairing the best translation into the target dialect against the best translation into a different dialect — improves GEMBA scores beyond the best SFT checkpoint.

The paper is organized as follows. Section 2 describes the related work. Section 3 describes the experimental setup, including base model selection, the GEMBA estimator, datasets, and finetuning procedures. Section 4 reports results for SFT, DPO, and the final segment-level ensemble.

Proceedings of the Eleventh Conference on Machine Translation, pages 1271–1283
October 28-29, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Arabic MT has evolved through the same broad paradigm shifts as the wider field. Early systems combined rule-based and phrase-based statistical approaches with dedicated morphological analysis to handle Arabic's complex inflectional morphology (Habash, 2010; Habash and Hu, 2009). The rise of Transformer-based NMT (Vaswani et al., 2017) rendered many of these hand-crafted components obsolete, yielding strong results on standard Arabic benchmarks (Almahairi et al., 2016; Durrani et al., 2016). More recently, large language models finetuned for Arabic have pushed quality further (Nagoudi et al., 2022a,b) – a broader comparison of current Arabic-capable LLMs is given in Table 1 and Figure 1. Dialectal Arabic MT poses additional challenges: parallel data are scarce, orthographic conventions are unstable, and dialects diverge substantially from MSA and from each other. Early approaches pivoted through MSA as an intermediate (Sajjad et al., 2013; Salloum and Habash, 2013); later work finetuned models directly on dialect data (Zbib et al., 2012; Nagoudi et al., 2021). The MADAR corpus provided the first broad-coverage dialectal resource (Bouamor et al., 2018), and Egyptian Arabic in particular has attracted attention as the most widely understood spoken variety (Darwish et al., 2021). IWSLT has included dialectal tracks (Tunisian, Levantine) since 2024, where LLMs have proved competitive (Ahmad et al., 2024; Abdulmumin et al., 2025; Ben Kheder et al., 2024; Jon et al., 2025).

The GEMBA framework addresses this by prompting an LLM to score translations on explicit criteria (Kocmi and Federmann, 2023a,b); we extend it with a dialect authenticity dimension and use it as the central quality signal throughout our pipeline. Preference optimization methods—DPO (Rafailov et al., 2023) and related contrastive objectives (CPO; Xu et al. (2024))—have recently been applied to MT to push finetuned models beyond what supervised training alone achieves.

3 Experimental setting

This section contains an overview of our experimental setting, including models, datasets and tools.

3.1 Base Models

We have run extensive evaluation of LLMs on translating from English into Arabic dialects, extending our previous work from Jon et al. (2026).

	CHRF		
Model	24++ sa	24++ eg	25 eg
EuroLLM-22B-Instruct	40.6	37.1	33.6
EuroLLM-9B-Instruct	39.9	35.7	31.0
Fanar-1-9B-Instruct	39.5	36.8	34.3
Fanar-2-27B-Instruct	42.7	39.4	35.4
Jais-2-70B-Chat	42.3	44.9	40.6
Jais-2-8B-Chat	39.6	42.8	37.4
Karnak-40B-v1.0	35.6	39.3	33.1
Karnak-6B-v1.0	41.1	41.1	37.6
Karnak-70B-LLAMA-v1.0	41.3	39.5	35.1
Llama-3.3-70B-Instruct	39.9	38.5	35.2
Mistral-Small-3.2-24B	38.9	35.7	32.2
Mistral-Small-4-119B	40.6	40.5	36.2
Nile-Chat-12B	38.5	42.7	38.0
Qwen3-14B	39.8	35.8	32.6
aya-expanse-32b	41.9	40.9	37.7
aya-expanse-8b	40.7	36.4	34.0
c4ai-command-r-08-2024	41.7	40.7	38.1
c4ai-command-r-v01	36.4	34.3	32.8
gemma-3-27b-it	42.5	41.3	39.3
gemma-4-12B-it	44.1	47.2	43.9
gemma-4-31B-it	44.2	47.5	44.5

Table 1: CHRF scores on WMT24++ Saudi and Egyptian Arabic (24++ sa and 24++ eg) and on WMT25 Egyptian Arabic test set (25 eg) for multiple LLMs.

We expanded the original list of compared models with newly published Gemma 4 (31B) and Karnak (6B, 40B and 70B) and the list of test sets by the WMT25 and WMT24++ English to Egyptian Arabic and Saudi Arabic test sets (Kocmi et al., 2025; Deutsch et al., 2025), while using the same inference setup and postprocessing. Table 1 and Figure 1 show ChrF scores on the WMT and MADAR datasets, respectively. Table 7 in Appendix A shows the corresponding BLEU scores on the WMT test sets. Note that the references are not secret, so they might have been used in the models' training data, especially in the case of the older MADAR test set.

Based on the results, we selected several models for finetuning:

- **Gemma-4-12B-it** – our primary model throughout the experiments. In automated metrics, it performs only slightly worse than the 31B and the number of parameters fits within the WMT constrained-track limit ($\leq$20B parameters).
- **Gemma-4-31B-it** – used for some of the experiments and GEMBA scoring, but not in the final submission.
- **Jais-2-8B-Chat** – a well-performing Arabic-specialized model.

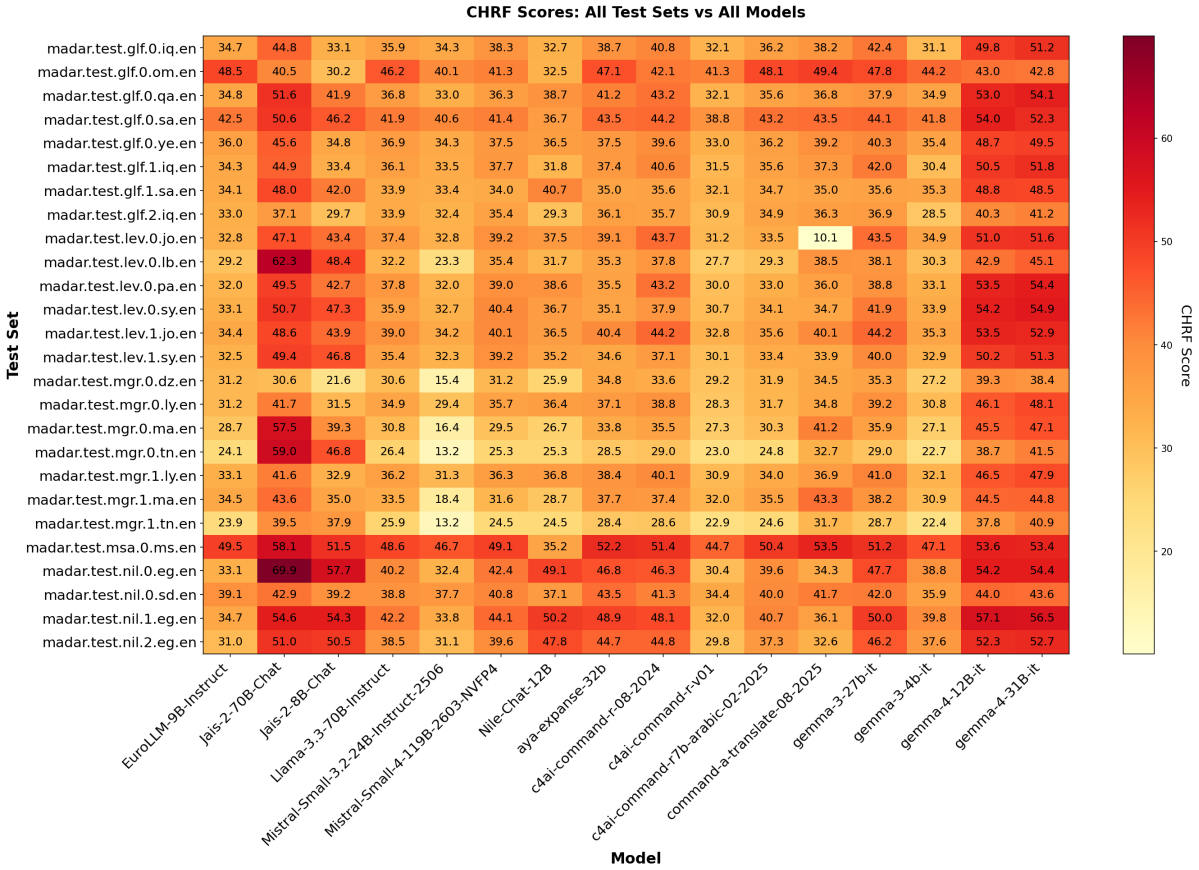

Figure 1: ChrF scores on MADAR test sets.

- **Aya-expanse-8B** – a lower-performing model included to assess how much finetuning can help.

3.2 GEMBA evaluation

Standard quality estimation (QE) models, like COMET-QE, are not suitable for dialectal Arabic, since it is not possible to indicate the target language/dialect. Thus, they have no way of taking the dialectalness into account when scoring. In multiple stages of our pipeline, we experimented with prompting multiple LLMs (method also called GEMBA) to score the translations along two dimensions:

- **Accuracy** — whether the translation faithfully conveys the meaning of the source.
- **Dialect authenticity** — whether the target text is in the desired Arabic variety.

We correlated the produced scores with the WMT25 human evaluation on the Egyptian Arabic translation. The results are shown in Table 2. We see that Gemma 4 GEMBA scores correlate quite well with the human scores, even though not for all the systems, suggesting that some types of errors made by these systems might go unnoticed. We chose to use Gemma 4 as the main source of the GEMBA scores in the following experiments, even though we note that our base models are also Gemma 4 and scoring their outputs by themselves is not optimal, for example in the DPO stage.

Table 3 reports GEMBA accuracy, dialect authenticity, and combined scores for a selection of WMT25 competition systems alongside our unmodified base models, evaluated by two judge sizes (Gemma-4-12B-it and Gemma-4-31B-it).

The split into accuracy and dialect immediately clarifies the low human rankings at the bottom of the WMT25 field: AyaExpanse-8B scores 57 on accuracy but only 1–3 on dialect authenticity under both judges, and Algharb shows the same pattern (accuracy 63–68, dialect 0.7–2.2). These systems translate into Modern Standard Arabic rather than Egyptian Arabic, which human raters penalize heavily — and GEMBA captures this correctly. Aya-expanse-8b re-evaluated with our pipeline on the same 199 WMT25 test segments, yields 58.7 / 3.1 / 30.9 (G-12B) and 56.8 / 1.4 / 29.1 (G-31B),

System	Hum.	Gemma-4-12B			Gemma-4-31B			Gemma-3-27B			Aya-Exp-8B			Jais-2-8B			Jais-2-70B			GPT-5.4-nano		
		Acc.	Dial.	Comb.	Acc.	Dial.	Comb.	Acc.	Dial.	Comb.	Acc.	Dial.	Comb.	Acc.	Dial.	Comb.	Acc.	Dial.	Comb.	Acc.	Dial.	Comb.
GPT-4.1	77.0	0.310	0.334	0.368	0.250	0.293	0.283	0.326	0.243	0.358	-0.071	0.003	-0.035	0.079	0.025	0.064	0.263	0.090	0.191	0.164	0.169	0.197
CommandA	74.0	0.260	0.677	0.609	0.309	0.650	0.611	0.214	0.115	0.190	0.211	0.128	0.206	0.066	-0.008	0.026	0.175	0.102	0.168	0.170	0.349	0.287
Gemini-2.5-Pro	60.6	0.587	0.974	0.942	0.718	0.973	0.955	0.163	0.493	0.378	-0.165	-0.126	-0.172	-0.015	-0.025	-0.026	0.147	0.519	0.492	0.153	0.675	0.583
DeepSeek-V3	56.8	0.326	0.877	0.819	0.330	0.838	0.782	-0.096	0.156	0.060	-0.182	-0.050	-0.134	0.103	-0.064	0.022	0.022	0.347	0.316	-0.063	0.506	0.352
Claude-4	55.7	0.599	0.879	0.855	0.638	0.860	0.841	0.026	0.369	0.233	-0.089	-0.078	-0.102	0.127	-0.064	0.002	0.117	0.366	0.346	0.051	0.659	0.556
IRB-MT	51.9	0.349	0.838	0.798	0.217	0.839	0.739	0.024	0.313	0.237	-0.141	-0.030	-0.089	0.038	0.011	0.030	0.075	0.314	0.290	-0.063	0.506	0.357
Mistral-Medium	36.0	0.565	0.916	0.884	0.567	0.932	0.887	-0.082	0.272	0.143	-0.142	0.001	-0.064	0.006	-0.112	-0.084	-0.041	0.364	0.305	-0.042	0.541	0.401
CommandA-MT	34.6	0.700	0.952	0.926	0.695	0.961	0.933	-0.165	0.316	0.170	-0.170	0.064	-0.018	-0.089	-0.202	-0.218	-0.207	0.282	0.096	-0.043	0.589	0.530
UvA-MT	29.0	0.553	0.838	0.820	0.546	0.893	0.868	-0.164	0.220	0.062	-0.164	0.111	0.015	0.038	-0.083	-0.037	-0.127	0.225	0.166	-0.076	0.507	0.443
CommandR7B	3.7	0.268	0.798	0.636	0.239	0.841	0.637	-0.087	-0.055	-0.079	-0.110	0.083	0.011	-0.030	0.090	0.046	-0.164	0.065	-0.167	-0.054	0.251	0.164
GemTrans	3.7	0.129	0.683	0.423	0.152	0.734	0.475	-0.211	-0.120	-0.185	-0.063	0.118	0.059	0.092	0.107	0.131	-0.134	0.076	0.030	-0.144	0.118	0.064
Algharb	3.2	0.092	0.624	0.289	0.157	0.609	0.355	-0.173	-0.088	-0.131	0.053	0.083	0.090	-0.006	0.042	0.027	-0.185	0.010	-0.165	-0.064	0.040	0.021
Shy	3.2	0.179	0.753	0.507	0.155	0.777	0.525	-0.071	0.068	0.026	-0.070	0.045	0.007	0.088	0.113	0.132	0.127	0.020	0.052	0.039	0.092	0.095
AyaExpanse-8B	2.0	0.114	0.691	0.350	0.123	0.816	0.425	-0.125	-0.155	-0.161	-0.065	0.030	0.001	-0.034	0.063	0.026	-0.049	0.039	-0.070	0.004	0.172	0.133
ONLINE-B	1.7	0.118	0.617	0.341	0.072	0.672	0.323	-0.126	-0.215	-0.207	-0.071	0.013	-0.004	0.064	0.063	0.089	0.001	-0.038	-0.008	-0.070	0.159	0.115
Yolu	1.4	0.091	0.545	0.286	0.137	0.754	0.388	-0.037	-0.209	-0.170	0.142	0.184	0.207	-0.020	0.075	0.037	0.056	0.087	0.096	0.031	0.133	0.116
Wenyiil	1.4	0.125	0.735	0.353	0.114	0.793	0.386	-0.029	-0.118	-0.104	0.147	0.108	0.149	0.059	0.073	0.085	0.010	0.053	0.050	0.012	0.012	0.014
SRPOL	0.9	0.102	0.701	0.326	0.109	0.836	0.402	0.006	-0.159	-0.119	0.160	0.165	0.202	0.056	0.050	0.070	0.160	0.102	0.134	0.112	0.128	0.151
Overall	27.6	0.617	0.948	0.910	0.643	0.952	0.916	-0.056	0.348	0.226	-0.253	0.031	-0.093	0.009	-0.042	-0.026	0.026	0.368	0.319	-0.017	0.551	0.465

Table 2: Pearson r between GEMBA and WMT25 human scores (English→Egyptian Arabic) by WMT25 system and GEMBA judge model. Acc. = accuracy, Dial. = dialect authenticity, Comb. = combined. Human scores are MQM-style DA on a 0–100 scale; systems sorted by average human score (descending).

compared to 56.8 / 2.5 / 29.7 and 55.4 / 0.9 / 28.2 in the WMT25 scoring. The small discrepancy is expected: mostly due to a different prompt and decoding parameters.

Overscoring by GEMBA The first form of overscoring concerns the absolute calibration of GEMBA scores against human direct assessment (DA). Although both GEMBA and human DA nominally use a 0–100 scale, they are not aligned. Even if we focus on the more correlated dialect score, it tends to overestimate the quality compared to human evaluators, especially at the high end (e.g. 77 human vs 94.5 dialect GEMBA for GPT-4.1).

The second form of overscoring, which we hypothesize, is towards the GEMBA evaluator model's own outputs (or outputs of the models from the same family). It is probable that the quality of Gemma-produced translations is overestimated by a Gemma-based GEMBA evaluator. As we do not have human evaluations available for the Gemma models we use, we could not test this hypothesis. However, we still used Gemma 4 models both for translating and quality estimation, keeping this risk in mind, since no other models came close to it in terms of correlation with human judgment.

3.3 Tools

Training All finetuning was implemented with the HuggingFace `transformers` library (Wolf et al., 2020) and the `TRL` library (von Werra et al., 2020) for SFT and DPO trainers. LoRA adapters were managed via `PEFT` (Mangrulkar et al., 2022). 4-bit QLoRA quantisation used `bitsandbytes` (Dettmers et al., 2023).

Inference Translation and GEMBA scoring were performed with `vLLM` (Kwon et al., 2023).

Evaluation ChrF (Popović, 2015) and BLEU (Papineni et al., 2002) were computed with `SacreBLEU` (Post, 2018). GEMBA scoring used our own implementation on top of vLLM-served judge models. Prompts are given in Appendix B.

3.4 Datasets

We have experimented with multiple datasets, listed in Table 4. We created the WMT23 dataset by translating all of the source segments of WMT23 dev sets using gpt-5.4-nano, Gemma-4-{12,31}B-it and Jais-2-{8,70}B-Chat into 21 Arabic dialects, which we name by the countries in which they are used: Egyptian Arabic (eg), Saudi Arabic (sa), Jordanian Arabic (jo), Palestinian Arabic (ps), Lebanese Arabic (lb), Syrian Arabic (sy), Moroccan Arabic (ma), Mauritanian Arabic (mr), Tunisian Arabic (tn), Algerian Arabic (dz), Libyan Arabic (ly), Iraqi Arabic (iq), Sudanese Arabic (sd), Omani Arabic (om), Yemeni Arabic (ye), Emirati Arabic (ae), Bahraini Arabic (bh), Kuwaiti Arabic (kw), Qatari Arabic (qa), Chadian Arabic (td) and Modern Standard Arabic (msa). We note that this is not a perfect approach to distinguish the dialects, since many dialects are used across multiple countries and multiple dialects can be used in a single country, but we have chosen this approach to align with the MADAR dev/test set we used for evaluation. We then scored all the translations with GEMBA with Gemma-4-31B-it model. These scores were used both for filtering the dataset for SFT and to choose the preferred translations for the preference opti-

System	Hum.	Gemma-4-12B-it			Gemma-4-31B-it		
		Acc.	Dial.	Comb.	Acc.	Dial.	Comb.
WMT25 Egyptian Arabic competition systems (sorted by human score)							
GPT-4.1	77.0	93.2	92.1	92.6	96.3	94.5	95.4
CommandA	74.0	90.7	86.7	88.7	94.6	85.7	90.2
Gemini-2.5-Pro	60.6	88.9	68.0	78.4	90.2	68.2	79.2
Mistral-Medium	36.0	78.2	45.5	61.9	77.4	36.2	56.8
CommandA-MT	34.6	75.5	41.5	58.5	78.2	38.6	58.4
AyaExpanse-8B	2.0	56.8	2.5	29.7	55.4	0.9	28.2
Algharb	3.2	63.0	2.2	32.6	67.9	0.7	34.3
Our base models on WMT25 Egyptian Arabic test set (no finetuning)							
Gemma-4-12B-it	—	95.4	93.0	94.2	96.9	93.7	95.3
Jais-2-8B-Chat	—	77.5	74.2	75.9	74.3	64.0	69.2
Aya-expanse-8B	—	58.7	3.1	30.9	55.0	1.2	28.1

Table 3: GEMBA accuracy (Acc.), dialect authenticity (Dial.), and combined score (Comb., 0–100) for selected WMT25 Egyptian Arabic competition systems and our unmodified base models, under two judge sizes. Human scores (Hum.) are MQM-style DA averages from the WMT25 evaluation.

mization methods. We call the filtered SFT dataset wmt23-gemba and the PO dataset wmt23-po.

Our backtranslated HPLT dataset is based on version 3.0 of originally a monolingual dataset compiled by the HPLT project[1]. It contains fine-grained language identification, differentiating between Arabic dialects (Levantine, Tunisian, Egyptian, Moroccan, Najdi and MSA). We backtranslated the texts into English using Gemma-4-31B-it to obtain a synthetic parallel dataset.

Alexandria conversational dataset (`alexandria`). The UBC-NLP/Alexandria dataset (EL Mekki et al., 2026) provides multi-turn bilingual conversations (English ↔ 13 Arabic dialectal varieties) covering Egypt, Jordan, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Saudi Arabia, Sudan, Syria, Tunisia, and Yemen. We used the full training split.

Other datasets listed in Table 4 are either parallel or monolingual resources found on HuggingFace.

For evaluation, we use WMT25 test set (Kocmi et al., 2025), Alexandria dev split and MADAR (Sajjad et al., 2020) dev and test sets.

3.5 Finetuning

SFT We have finetuned the pretrained models using QLoRA. We compared Aya-expanse-8B, Gemma-4-12B-it and Jais-2-8B-Chat as the base pretrained models and we finetuned them on different combinations of all datasets. Target dialect sets ranged from Egyptian Arabic only to 21 Arabic varieties. As the sweep narrowed, Gemma-4-12B-it and Jais-2-8B-Chat became the primary base models, with training data converging on GEMBA-filtered WMT23 synth data with and without Alexandria (other datasets did not bring any increases in ChrF scores on the WMT25 test set). Across all models we swept LoRA rank $r \in \{4, 8, 16, 32, 64, 128\}$, LoRA $\alpha = 2r$ and learning rates $\{$1e-5, 2e-5, 4e-5, 5e-5, 6e-5, 6.36e-5$\}$. Training ran for up to 8,000 steps in early broad runs and was capped at 1,000 steps in focused later runs.

DPO We generated preference pairs from the same pool of WMT23 hypotheses used for SFT, i.e., for each of the 22,518 WMT23 source sentences, multiple translations were produced across 21 Arabic dialect varieties, each scored by GEMBA (Gemma-4-31B-it judge) along two dimensions: accuracy and dialect authenticity. We used the mean of the two scores as a single combined quality signal.

We constructed two complementary pair types. **Within-dialect pairs** contrast translations of the same source into the same target dialect produced by different translation models: the higher-scoring hypothesis is *chosen* and any lower-scoring hypothesis is *rejected*, provided the chosen score meets a minimum threshold (combined GEMBA ≥ 95) and the score gap is positive. **Cross-dialect pairs** use the best-scoring translation into dialect A as the chosen side and the best-scoring translation of the same source into a different dialect B as the rejected side, where the prompt requested dialect A. This gives the model a direct signal that producing the wrong variety is undesirable, even if that variety would otherwise score well on accuracy. To avoid spurious negative signal, we ex-

[1] https://hplt-project.org/datasets/v3.0

Dataset (HuggingFace ID)	Dialects	Size
Parallel corpora (SFT)		
UBC-NLP/alexandria	13 varieties[b]	∼70K turns
IbrahimAmin/arz-en-parallel-corpus	Egyptian	24K segs
HeshamHaroon/ArzEn-MultiGenre	Egyptian	25K segs
HeshamHaroon/Egyptian_English_parallel	Egyptian	874 segs
Omar-youssef/English-Egyptian-Arabic-Translation-Pairs	Egyptian	8K segs
Monolingual, back-translated (SFT)		
yrrhall/arabic-dialect-corpus	MSA, Egyptian, Saudi	150K segs
Omartificial-Intelligence-Space/FineWeb2-Egyptian-Arabic	Egyptian	50K segs
Omar-youssef/QA_LAW_Egyptian_dataset	Egyptian	3.7K segs
HPLT 3.0 (web crawl, back-translated)	Multiple	—
Monolingual, forward-translated (SFT)		
WMT23 translations[a]	21 varieties	∼200K segs
Preference pairs (DPO/CPO)		
WMT23 pref. pairs[a]	21 varieties	536k pairs

Table 4: All training data sources used across SFT and DPO experiments. Segment counts reflect training examples (source–hypothesis pairs) after dialect filtering; sizes vary by dialect subset used in each run.
[a] Derived from WMT23 English–Arabic system outputs translated into 21 Arabic varieties using multiple MT systems (22,518 source segments per dialect).
[b] EG, JO, LB, LY, MA, MR, OM, PS, SA, SD, SY, TN, YE.

cluded cross-dialect pairs where the two dialects belong to the same equivalence group (Gulf cluster: ae/bh/kw/om/qa/sa; Levantine: jo/ps; Maghrebi: dz/ly/ma/mr/td). The resulting dataset contains 535,688 pairs in total: 344,513 within-dialect and 191,175 cross-dialect.

We applied DPO either starting from the base model weights or from an SFT checkpoint in the final experiments. Training used either a CPO loss or a composite loss combining the standard DPO sigmoid objective (weight 0.8), a BCO pair loss (weight 0.2), and an SFT cross-entropy loss on the chosen translations (weight 1.0). We set $\beta = 0.1$, trained for 3 epochs, and used a batch size of 64 (1 example, 64 gradient accumulation steps), maximum sequence length 512 tokens, LoRA $r \in \{8, 64, 128\}$, LoRA $\alpha = 2r$ and learning rates $\in \{3\text{e-}5,\ 6\text{e-}5\}$. Checkpoints were saved and evaluated every 25 steps against the WMT25 Egyptian Arabic held-out set via ChrF.

4 Results

This section describes the results obtained in our experiments.

4.1 Supervised finetuning

We ran 123 SFT experiments spanning three base models, six LoRA ranks, six learning rates, and multiple data combinations. We present our findings in the following paragraphs. ChrF and GEMBA scores on the WMT25 test set for best or interesting training runs are shown in Table 5, the corresponding WMT26 GEMBA scores are in Table 9 in the Appendix.

Base model Gemma-4-12B-it was the winner across all configurations. It achieved average combined GEMBA of 94.23 and ChrF of 42.65 on the WMT25 validation set, compared to 75.88 GEMBA and 36.63 chrF for Jais-2-8B-Chat and 30.88/32.67 Aya-expanse-8B.

SFT doesn't improve metrics of a good base model A notable result is that the unmodified Gemma-4-12B-it base model already scores 42.65 ChrF and 94.23 GEMBA on the WMT25 validation set – higher than most finetuned checkpoints. We actually used 4-bit quantized version of the model for the finetuning, so the starting scores were 42.1 and 93.54, but we consider it fairer to compare the results with the full precision baseline.

ChrF and GEMBA scores of finetuned models rarely exceeded the base model (or by a very small margin compared to 4-bit quantized version of the base model), suggesting that Gemma-4-12B-it already produces high-quality Egyptian Arabic and that even our best SFT configs do not help with the translation quality. Also, we are computing the GEMBA scores with the Gemma-4-12B-it model itself, so the scores might be biased towards the base, non-finetuned model. Additionally, the finetuning is done on WMT23-GEMBA dataset, where most of the selected target translations are already pro-

r	lr	Data	@step	ChrF	GEMBA
Base model — Gemma-4-12B-it (no finetuning)					
—	—	—	—	42.65	94.23
Base model — Jais-2-8B-Chat (no finetuning, not used in the final submission)					
—	—	—	—	36.63	75.88
Base model — Aya-expanse-8B (no finetuning, not used in the final submission)					
–	—	—	—	32.67	30.88
SFT— Aya-expanse-8B (best SFT checkpoint, not used in the final submission)					
128	5e-5	G+A (eg)	800	39.50	79.14
SFT — Gemma-4-12B-it					
128	6e-5	G (eg)	400	42.21	94.07
128	6e-5	G (eg)	700	42.22	93.56
128	6e-5	G+A (eg)	500	42.05	93.75
128	6e-5	G+A (eg)	600	42.40	94.18
128	6e-5	G+A (eg-tn-msa)	300	42.29	93.54
128	6e-5	G+A (eg-tn-msa)	600	42.64	94.06
128	6e-5	G+A (eg-tn-msa)	1000	42.46	93.71
128	6e-5	G+A (eg-tn-msa)*	400	42.26	93.48
128	6e-5	G+A (eg-tn-msa)*	500	42.39	93.98
32	2e-5	G (eg-tn-msa)	50	42.63	94.34
32	5e-6	G (eg-tn-msa)	100	42.66	94.08
32	5e-6	G (eg-tn-msa)	300	42.69	94.24
SFT — Jais-2-8B-Chat					
128	6e-5	G+A (eg)	200	39.65	88.43
128	6e-5	G+A (eg-tn-msa)	1000	40.50	87.54
128	6e-5	G+A (eg-tn-msa)*	500	40.28	88.53
128	6e-5	G+A (eg-tn-msa)*	800	40.52	87.99
PO — DPO from base, Gemma-4-12B-it, within + cross-dialect, eg-tn-msa					
128	3e-5	`wmt23-po`	25	42.25	94.94
128	3e-5	`wmt23-po`	50	42.90	94.92
128	3e-5†	`wmt23-po`	25	42.21	94.63
128	3e-5†	`wmt23-po`	50	42.76	94.29
128	6e-5	`wmt23-po`	25	41.71	94.48
128	6e-5	`wmt23-po`	50	42.70	93.92
128	3e-5	`wmt23-po`	25	42.37	94.90
128	3e-5	`wmt23-po`	50	42.75	94.82
128	6e-5	`wmt23-po`	25	41.74	94.55
128	6e-5	`wmt23-po`	75	43.25	94.39
128	6e-5‡	`wmt23-po`	50	43.64	93.29
128	6e-5‡	`wmt23-po`	75	42.54	93.81
PO — DPO from SFT@500, Gemma-4-12B-it, `wmt23-po`*, min-sc* $= 90$					
128	5e-5	`wmt23-po`	20	41.84	93.95
128	5e-5	`wmt23-po`	20	41.75	94.00
Final ensemble — segment-level GEMBA selection across all 31 checkpoints					
—	—	—	—	42.62	**97.41**

Table 5: WMT25 Egyptian Arabic test-set scores (Gemma-4-12B-it judge; 199 segments) for all 31 checkpoints in the final pool and the segment-level ensemble. GEMBA is $\frac{1}{2}(\text{acc} + \text{dial})$. G = GEMBA-filtered WMT23; A = Alexandria. * pairs of runs with identical hyperparameters, different seeds; † ‡ same for DPO runs.

duced by the Gemma-4-31B-it or Gemma-4-12B-it, leaving too few new translations for the model to learn from. For Jais-2-8B-Chat, the improvements in both metrics after SFT are more clear. Table 9 in Appendix C shows that there are some gains in GEMBA scores after SFT on the WMT26 test set even for the Gemma model.

Training data Most datasets led to a decrease in automated metrics. The best-performing dataset was our synthetic WMT23-GEMBA, filtered to only contain examples with GEMBA score $>$ 95. Adding Alexandria conversational data to the GEMBA-filtered WMT23 translations (`G+A`) consistently improved dialect authenticity scores over WMT23-GEMBA-only training (`G`), with the gains most visible on ChrF. Including all 21 Arabic dialect targets in training (`eg-tn-msa` and broader sets) generally matched or exceeded Egyptian-only

Model / training category	Segments	%
Gemma-4-12B-it (no finetuning)	200	30.9
SFT — wmt23-gemba, no alx	152	23.5
SFT — wmt23-gemba + alx (Gemma)	49	7.6
SFT — wmt23-gemba + alx (Jais)	27	4.2
DPO — wmt23-pairs	197	30.4
Alexandria SFT → DPO	23	3.5
Total	**648**	**100.0**

Table 6: Segment-level contribution of each model category in the final ensemble submission. For each of the 648 test segments (Egyptian Arabic arz/arz_Arab, Tunisian aeb, MSA ar_AR), the hypothesis with the highest combined GEMBA score (Gemma-4-12B-it judge) was selected from 31 candidate checkpoints. Alx means Alexandria dataset.

training, likely because broader dialect coverage acts as a regularizer.

Hyperparameters LoRA rank $r = 128$ was the strongest setting for most configurations, though $r = 32$ with a lower learning rate (2e-5) achieved the highest finetuned-checkpoint GEMBA score (94.34). The best-performing learning rate was 6e-5 for $r = 128$ runs. Checkpoints typically peaked early (steps 50–700) and degraded or plateaued thereafter, so the 1,000-step training budget was sufficient.

4.2 Preference optimization

Aside from the hyperparameters, we varied the preference pairs dataset for the training. We tried filtering it by GEMBA score thresholds of 80, 90 and 95 for the positive examples. We have also tried using only intra-dialect pairs vs. cross-dialect pairs. We compared the DPO and CPO loss functions.

DPO training on high-quality preference pairs (chosen GEMBA ≥ 95, including both within-dialect and cross-dialect pairs) improved the combined GEMBA score by ≈ 0.7 points over the best SFT checkpoint on the WMT25 test set and by ≈ 0.3 points on WMT26. The best DPO configuration used lr = 3e-5, $\beta = 0.1$.

4.3 Final submission

For the final submission, we selected 20 runs with the best ChrF scores on the dev set (first 150 sentences of the WMT25 test set) across SFT and PO finetuned models. From these 20 runs, we then selected the top 30 checkpoints based on the GEMBA score on the WMT25 test set, with at most 2 checkpoints per each run (model configuration) selected. We translated the WMT26 test set with each selected checkpoint and we also added translations by the original base model to the pool. We computed the GEMBA score on the WMT26 test set using Gemma-4-12B. For each test example, we selected the best-scoring translation. The per-model type breakdown is shown in Table 6. Almost all of the translation models were Gemma-4-12B with small LoRA adapters on top of them, so even though the submission is not considered constrained by the official rules, the total number of different parameters involved is under the 20B limit (same base 12B for translation and scoring and a small number of LoRA adapter parameters on top of them). The only exception is 27 (4.2%) sentences that were translated by a finetuned Jais 8B model. 30.9% of translations were chosen from the base model without any finetuning, however, this exact model was used for the GEMBA scoring for the selection, so preference for its own outputs could be expected. The per-model breakdown for the WMT25 test set ensemble is reported in Table 8 in the Appendix C.

4.4 Manual inspection

To complement the automatic evaluation, we manually inspected the first 50 translations and their Gemma-12B-it GEMBA assessments. The inspection suggests that, although the LLM is generally effective at identifying clear differences in translation quality, it can overlook errors that are subtle but noticeable to a native speaker, particularly those affecting fluency, dialect authenticity, lexical choice, and overly literal or word-for-word phrasing. Several translations received very high accuracy or dialect scores despite containing slightly unnatural Egyptian Arabic, MSA-influenced spelling, awkward constructions, minor mistranslations, or typographical errors. A particularly illustrative case is the first inspected example, "You could tell those fuckers where to get off." The explicit insult "fuckers" is not faithfully represented in the translation, yet the GEMBA reasoning effectively treats the Arabic expression as accounting for both the insult and the idiom, leading to a high assessment despite the loss of this source-side detail. We also observed cases of word-by-word translations, where the resulting sentence was unnatural or even difficult to interpret as a whole, as well as overly literal translations that GEMBA scored close to perfectly. Overall, the manual inspection indicates that our Gemma-12B-it GEMBA assessment captures broad semantic adequacy and dialect differences

well, but is less reliable for fine-grained errors in naturalness, register, idiomaticity, and exact meaning preservation that are readily apparent to a native speaker. It is important to note again that the translations were also produced by a finetuned Gemma-12B-it model, so it is expected that when the same model is used as a judge, some translation errors may be ignored. We plan to carry out a more extensive evaluation after the human assessment scores of all the competing systems are published.

5 Conclusions

We have compared multiple base models and finetuning methods and datasets with the goal of improving the translation quality on the WMT26 English to Egyptian Arabic task. We saw some increases in various automatic metrics, however, the possible real improvement in translation quality will become clearer after human evaluation. GEMBA with a Gemma-4-31B-it judge correlated strongly with WMT25 human scores ($r = 0.916$), validating its use as a development signal throughout the pipeline. DPO on high-quality cross-dialect preference pairs provided the largest single quality step beyond SFT, while segment-level ensemble selection over 31 checkpoints further boosted scores by exploiting complementary strengths of individual models. The effect on real, human-perceived translation quality remains unknown.

Limitations

The limitations of our paper are shared across all LLM-based research. The evaluation of our methods is limited by the fact that we use pretrained LLMs, so it is possible that our test data were part of their training data. A more specific concern is that we use the Gemma-4 family both as the primary translation model and as the GEMBA judge. A judge model may favor outputs from models in the same family, which could inflate GEMBA scores for our Gemma-based checkpoints and bias segment selection in the ensemble toward them.

Acknowledgements

This work was supported by the project TQ01000458 (EdUKate) financed by the Technology Agency of the Czech Republic (www.tacr.cz) within the Sigma 3 Programme. It has been also partially supported by the grant TQ28000003 (NCUI, Národní centrum umělé inteligence), and the grant CZ.02.01.01/00/23_020/0008518 ("Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím"). It has been also partially supported by by SVV project number 260 821.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John P. McCrae, and 25 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville. 2016. First result on arabic neural machine translation. *Preprint*, arXiv:1606.02680.

Mohamed Anwar, Abed Alhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Sadallah, Gurpreet Gosal, Gokulakrishnan Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah AlBarri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2026. Jais 2: A family of arabic-centric open large language models. *Preprint*, arXiv:2608.13580.

Waad Ben Kheder, Josef Jon, André Beyer, Abdel Messaoudi, Rabea Affan, Claude Barras, Maxim Tychonov, and Jean-Luc Gauvain. 2024. ALADAN at IWSLT24 low-resource Arabic dialectal speech translation task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 192–202, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The

MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Samhaa R. El-Beltagy, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Preprint*, arXiv:2011.12631.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2016. QCRI's machine translation systems for IWSLT'16. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Abdellah EL Mekki, Samar M. Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, Thikra Al-hibiri, Razan Saadie, Hamzah A. Alsayadi, Nadia Ghezaiel Hammouda, Alshima Mohammed Alkhazimi, Aya Hamod, Al-Yas Yaqoob Al-Ghafri, Wesam El-Sayed, Asila Ismail al Sharji, Mohamad Ballout, Anas Belfathi, Karim Ghaddar, Serry Sibaee, and 28 others. 2026. Alexandria: A multi-domain dialectal Arabic machine translation dataset for culturally inclusive and linguistically diverse LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32567–32592, San Diego, California, United States. Association for Computational Linguistics.

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece. Association for Computational Linguistics.

N.Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers.

Josef Jon, Waad Ben Kheder, André Beyer, Claude Barras, and Jean-Luc Gauvain. 2025. ALADAN at IWSLT25 low-resource Arabic dialectal speech translation task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 252–259, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Josef Jon, Rawan Bondok, and Ondřej Bojar. 2026. Current state of LLMs for Arabic dialectal machine translation. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 329–363, Rabat, Morocco. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022a. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022b. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English machine translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, Atlanta, Georgia. Association for Computational Linguistics.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

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.

A Base models' BLEU scores

Model	BLEU		
	24++ sa	24++ eg	25 eg
EuroLLM-22B-Instruct	9.5	7.3	5.3
EuroLLM-9B-Instruct	9.1	6.6	4.1
Fanar-1-9B-Instruct	8.5	7.1	5.3
Fanar-2-27B-Instruct	10.5	9.0	5.6
Jais-2-70B-Chat	10.9	14.2	9.9
Jais-2-8B-Chat	9.0	12.5	8.4
Karnak-40B-v1.0	7.9	10.2	6.7
Karnak-6B-v1.0	9.1	11.0	8.0
Karnak-70B-LLAMA-v1.0	9.9	9.4	6.2
Llama-3.3-70B-Instruct	8.7	8.1	5.9
Mistral-Small-3.2-24B	8.8	7.0	4.8
Mistral-Small-4-119B	9.3	10.6	7.4
Nile-Chat-12B	7.9	11.8	7.6
Qwen3-14B	8.3	6.2	4.1
aya-expanse-32b	9.8	10.2	7.3
aya-expanse-8b	8.7	6.1	4.6
c4ai-command-r-08-2024	9.8	10.1	7.9
c4ai-command-r-v01	6.8	5.1	3.9
gemma-3-27b-it	10.1	9.9	8.6
gemma-4-12B-it	11.5	15.4	12.9
gemma-4-31B-it	11.8	15.5	13.6

Table 7: BLEU scores on WMT24++ Saudi and Egyptian Arabic (24++ sa and 24++eg) and on WMT25 Egyptian Arabic test set (25 eg) for multiple LLMs.

Table 7 shows BLEU scores on WMT Arabic test sets for multiple LLMs.

B Prompts

We have used the following prompt for all translations:

```
<bos><|turn>system You are a professional, very precise translator and a native speaker. Translate inputs based on the instructions and always print out only the text of the best possible translation, no explanations. Keep the same formatting (e.g. markup, lines, spacing) as the original. Do not translate untranslatable parts of the input (URLs, code and similar).<turn|> <|turn>user Translate the following text into {dialect} Arabic, only print out the translation. Do not add any explanations: {text} <turn|>
```

We have used the following prompts for the GEMBA scoring: Accuracy:

```
Score the following translation from {source_lang} to {target_lang} on a continuous scale from 0 to 100 for translation accuracy only (meaning preservation and grammar). You can briefly reason about the errors and describe them, but make sure to output a Score: and a single number as the final score at the end of the output.
{source_lang} source: {source_seg}
{target_lang} translation: {target_seg}
Evaluation:
```

Dialect:

```
Score how authentically the following translation uses {target_lang} (appropriate dialect, register, and locale conventions) on a scale from 0 to 100, where 0 means not authentic / wrong variety and 100 means fully authentic {target_lang}. If a dialect is specified and the translation is entirely in a different dialect, score the translation 0. You can briefly reason about the errors and describe them, but make sure to output a Score: and a single number as the final score at the end of the output.
{source_lang} source: {source_seg}
{target_lang} translation: {target_seg}
Evaluation:
```

C Additional Results

Table 8 reports the per-model breakdown of segment selection in the WMT25 Egyptian Arabic ensemble (199 segments, same 31-checkpoint pool as the WMT26 submission). Table 9 reports GEMBA scores on the WMT26 test set broken down by dialect for all 31 checkpoints.

Model / training category	**Segments**	**%**
Gemma-4-12B-it (no finetuning)	48	24.1
SFT — wmt23-gemba	39	19.6
SFT — wmt23-gemba + alx (Gemma)	23	11.6
SFT — wmt23-gemba + alx (Jais)	7	3.5
DPO from base	75	37.7
DPO from SFT	7	3.5
Total	**199**	**100.0**

Table 8: Segment-level contribution of each model category in the WMT25 Egyptian Arabic ensemble. For each of the 199 test segments the hypothesis with the highest combined GEMBA score (Gemma-4-12B-it judge) was selected from 31 candidate checkpoints.

				GEMBA avg (0–100)				
r	**lr**	**Data**	**@step**	**aeb**	**ar_AR**	**arz**	**arz_Arab**	**Avg**
Base model — Gemma-4-12B-it (no finetuning)								
—	—	—	—	86.57	97.26	94.08	93.64	92.89
Base model — Jais-2-8B-Chat (no finetuning, not used in the final submission)								
—	—	—	—	69.23	92.36	85.97	78.09	81.41
SFT — Gemma-4-12B-it								
128	6e-5	G (eg)	400	85.37	97.58	95.30	95.28	93.38
128	6e-5	G (eg)	700	86.29	97.51	94.42	95.17	93.35
128	6e-5	G+A (eg)	500	85.53	97.54	94.91	95.61	93.40
128	6e-5	G+A (eg)	600	85.91	97.24	94.59	95.41	93.29
128	6e-5	G+A (eg-tn-msa)	300	86.10	97.67	94.16	94.47	93.10
128	6e-5	G+A (eg-tn-msa)	600	88.93	97.15	94.15	94.80	93.76
128	6e-5	G+A (eg-tn-msa)	1000	88.58	97.52	94.95	95.00	94.01
128	6e-5	G+A (eg-tn-msa)*	400	88.29	96.98	94.37	95.27	93.73
128	6e-5	G+A (eg-tn-msa)*	500	89.78	97.15	93.28	95.91	94.03
32	2e-5	G (eg-tn-msa)	50	87.54	97.26	94.13	94.06	93.25
32	5e-6	G (eg-tn-msa)	100	86.20	97.13	94.28	94.16	92.94
32	5e-6	G (eg-tn-msa)	300	86.47	97.61	94.26	94.61	93.24
SFT — Jais-2-8B-Chat								
128	6e-5	G+A (eg)	200	81.84	94.80	92.54	85.19	88.59
128	6e-5	G+A (eg-tn-msa)	1000	87.04	94.44	92.30	88.95	90.68
128	6e-5	G+A (eg-tn-msa)*	500	87.72	95.23	92.78	87.87	90.90
128	6e-5	G+A (eg-tn-msa)*	800	86.90	94.45	92.74	88.47	90.64
PO — DPO from base, Gemma-4-12B-it, within + cross-dialect, eg-tn-msa								
128	3e-5	`wmt23-po`	25	88.45	97.66	95.18	95.91	94.30
128	3e-5	`wmt23-po`	50	87.67	97.29	94.66	95.90	93.88
128	3e-5†	`wmt23-po`	25	87.59	97.50	94.90	95.49	93.87
128	3e-5†	`wmt23-po`	50	85.40	96.08	92.32	96.26	92.51
128	6e-5	`wmt23-po`	25	85.87	96.89	94.44	96.21	93.35
128	6e-5	`wmt23-po`	50	79.67	89.12	86.60	96.03	87.85
128	3e-5	`wmt23-po`	25	86.34	97.51	94.57	95.65	93.52
128	3e-5	`wmt23-po`	50	87.90	97.35	93.59	96.47	93.83
128	6e-5	`wmt23-po`	25	85.69	97.45	94.53	95.48	93.29
128	6e-5	`wmt23-po`	75	82.39	90.60	89.11	95.74	89.46
128	6e-5‡	`wmt23-po`	50	79.79	93.72	89.98	95.61	89.77
128	6e-5‡	`wmt23-po`	75	71.92	86.36	82.74	95.23	84.06
PO — DPO from SFT@500, Gemma-4-12B-it								
128	5e-5	`wmt23-po`	20	86.14	97.25	94.30	94.85	93.14
128	5e-5	`wmt23-po`	20	85.83	96.27	94.42	94.36	92.72
Final submission — segment-level GEMBA selection across all 31 checkpoints								
—				97.00	99.70	98.69	98.16	**98.39**

Table 9: WMT26 GEMBA scores (Gemma-4-12B-it judge) for all checkpoints in the final pool and the combined submission, broken down by target dialect. G = GEMBA-filtered WMT23 translations; A = Alexandria. * pairs of Gemma SFT runs with identical hyperparameters, different seeds. † ‡ same for DPO runs.