

OpenBracket at WMT 2026: COMET-MBR System Combination over Arabic-Native and Multilingual Models for Low-Resource Arabic–English Translation

Abrar Alotaibi

College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, Dammam, KSA
amotaibi@iau.edu.sa

Abstract

We describe OpenBracket’s submission to the WMT 2026 shared task on Low-Resource Arabic–Asian Language Translation, covering Sub-Task 1A (English→Arabic) and Sub-Task 2A (Arabic→English). We build a pool of individually strong translators by fine-tuning a multilingual sequence-to-sequence model (NLLB-3.3B), an Arabic-capable translation LLM (X-ALMA, with an additional contrastive-preference stage), an Arabic-native instruction LLM (ALLaM-7B), and a bilingual baseline (OPUS-MT). Our primary system combines the best-1 output of each candidate through system-combination Minimum Bayes Risk (MBR) decoding with COMET-22 as the utility function. It yields the best COMET in both directions and, under paired bootstrap, no single system significantly outperforms it on BLEU, chrF, or COMET. Two contrastive systems expose complementary single-model strengths: an ALLaM-7B fine-tune attains the highest BLEU in both directions, and sampling-based MBR over NLLB-3.3B the highest COMET among individual systems. All systems use publicly available pretrained checkpoints and are trained on the official task data. On the official blind test set, our primary ranks first in Arabic→English, winning BLEU, chrF, and TER, and obtains the highest COMET on the English→Arabic primary leaderboard. We release our training and decoding scripts.

1 Introduction

Arabic is among the most widely spoken languages in the world, yet many Arabic-centric language pairs remain under-served by parallel data, which limits the quality of machine translation (MT) systems for these directions. The WMT 2026 shared task on Low-Resource Arabic–Asian Language Translation (Laskar et al., 2026) targets exactly this gap, with forward and backward directions between Arabic and a set of Asian languages. We participate in the two English-paired directions:

Sub-Task 1A (English→Arabic) and Sub-Task 2A (Arabic→English).

Our central design choice is to treat the low-resource setting as a *model-diversity* problem. Rather than betting on a single architecture, we fine-tune several pretrained models that bring different inductive biases to Arabic: a massively multilingual encoder-decoder (NLLB Team et al., 2022), a translation-specialized decoder-only LLM (Xu et al., 2025), an Arabic-native instruction LLM (Bari et al., 2025), and a compact bilingual baseline (Tiedemann and Thottingal, 2020). We then reconcile their outputs with Minimum Bayes Risk decoding under a neural utility metric (Freitag et al., 2022), which lets a quality estimator rather than model probability decide the final translation.

We make three observations on the development set. (i) A COMET-MBR combination over the deduplicated candidate pool attains the best COMET in both directions and is never significantly worse than any single system on any metric (with one exception on TER), so we use it as our primary submission. (ii) Among the individual systems, an Arabic-native LLM (ALLaM-7B), once fine-tuned on the official data, delivers the highest BLEU in both directions. (iii) Sampling-based MBR (hereafter sampling-MBR) over NLLB-3.3B delivers the highest COMET among the individual systems. The two contrastive systems were chosen to surface findings (ii) and (iii); the combination matches each of them on the metric where it is strongest and beats it on the metrics where it is not. On the official blind test set, the primary ranks first in Arabic→English and obtains the best COMET on the English→Arabic primary leaderboard (Section 7).

2 Task and Data

The shared task (Laskar et al., 2026) provides official training, development, and DevTest splits,

released to registered participants, and a held-out challenge test set for the evaluation cycle. The data is a single bidirectional Arabic–English parallel corpus comprising 21,000 training, 1,000 development, and 500 DevTest sentence pairs, shared by both sub-tasks. Manual inspection indicates general-domain, news-style content (politics, economics, sports, and cultural commentary). We use the official parallel data only for training and use no monolingual or external parallel data; our systems are therefore constrained in the sense of the task guidelines, while drawing on publicly available pretrained checkpoints. With only 21k pairs, the setting is low-resource by any standard, which motivates our reliance on strong pretrained models and quality-aware decoding rather than training large models from scratch.

Preprocessing. For the Arabic→English direction we normalize the Arabic source before training and inference: Alef variants are unified, and tatweel and diacritics are removed. We otherwise keep the official tokenization and rely on each model’s native subword vocabulary.

3 System Overview

We submit three systems per direction: one primary and two contrastive, respecting the maximum of three submissions per language pair. The primary is a system-combination MBR ensemble; the contrastive systems are a single fine-tuned LLM and a sampling-based MBR system. Figure 1 summarizes how the components relate.

3.1 Primary: COMET-MBR System Combination

The primary system combines the outputs of the single-system translators through system-combination MBR with COMET-22 (Rei et al., 2022a) as the utility function. For each source sentence, we form a deduplicated pool of candidate translations and select the candidate that maximizes the expected COMET utility over the pool. Intuitively, MBR with a neural metric prefers the hypothesis that is, on average, most similar to the other strong hypotheses under a quality-aware similarity, rather than the single highest-probability decode (Eikema and Aziz, 2020; Freitag et al., 2022). The pool is the same for both directions and contains exactly one best-1 hypothesis from each of five fine-tuned systems: NLLB-3.3B FT, X-ALMA-FT, X-ALMA-CPO, ALLaM-FT, and OPUS-MT.

This pool is used both for the development-set evaluation in Tables 1 and 2 and for the submitted test primary, so the development scores are a faithful predictor of the submitted system. Neither contrastive system enters the pool: the sampling-MBR output (Contrastive 2) is not added to the pool, and ALLaM-FT participates only through its single greedy output, which is the same output submitted as Contrastive 1. Adding the sampling-MBR hypothesis to the pool is a natural extension that we leave for future work.

3.2 Contrastive 1: ALLaM-7B Fine-Tune

ALLaM-7B-Instruct is an Arabic-native, instruction-tuned LLM (Bari et al., 2025). We fine-tune it with QLoRA (Dettmers et al., 2023; Hu et al., 2022) on the official parallel data, separately per direction (Section 4.5), and decode greedily under the model’s chat template. This system attains the highest BLEU on the development set in both directions, ahead of every other system including the combination.

3.3 Contrastive 2: Sampling-MBR over NLLB-3.3B

We draw $K=32$ hypotheses per source from the fine-tuned NLLB-3.3B model with nucleus sampling (Holtzman et al., 2020) (temperature 0.8, top- p 0.95). We then apply a two-stage MBR procedure: CometKiwi (Rei et al., 2022b) is used as a reference-free quality-estimation pre-pruning step that keeps the top 16 hypotheses per source, after which COMET-22 MBR scoring selects the final output over the pruned pool. Among the individual (non-combination) systems, this attains the highest COMET on the development set in both directions; the system-combination primary matches it on COMET and improves substantially on BLEU, chrF, and TER (Section 6).

4 Training Pipeline

We train each component of our submission separately and combine their outputs only at decoding time. This section describes the publicly available base checkpoints we start from, the fine-tuning recipe for each model (NLLB-3.3B, the X-ALMA SFT and CPO stages, and ALLaM-7B), and the decoding and ensembling procedure that yields the primary and contrastive systems. All models are fine-tuned on the official parallel data only, with the preprocessing described in Section 2.

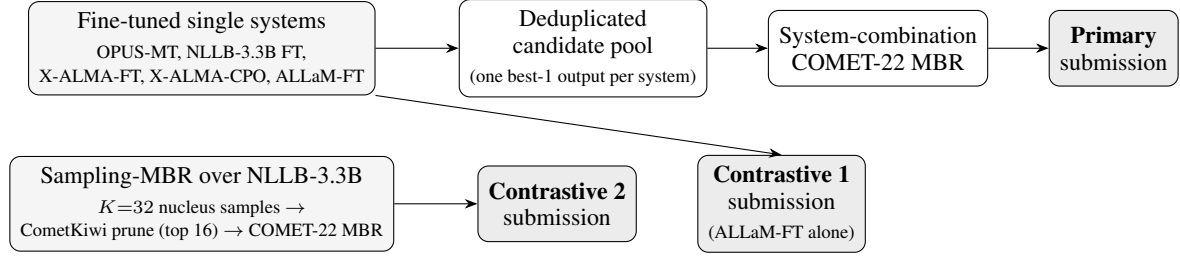


Figure 1: System overview. The five fine-tuned single systems each contribute one best-1 output to a deduplicated candidate pool, and system-combination MBR with a COMET-22 utility selects the primary translation. The two contrastive systems are stand-alone submissions and do *not* enter the primary pool: Contrastive 1 is the ALLaM-FT single system, and Contrastive 2 is sampling-MBR over NLLB-3.3B.

4.1 Base models

We initialize from publicly available checkpoints: NLLB-3.3B (NLLB Team et al., 2022); the X-ALMA-13B backbone with its Arabic language module (Xu et al., 2025), which we also report without fine-tuning as a zero-shot baseline (*X-ALMA pretrain* in the tables); ALLaM-7B-Instruct (Bari et al., 2025); and the bilingual OPUS-MT models for both directions (Tiedemann and Thottungal, 2020). All fine-tuning uses the Hugging Face Transformers stack (Wolf et al., 2020).

4.2 NLLB fine-tuning

NLLB-3.3B is fine-tuned with its layers sharded across GPUs under a per-device memory budget. We train in fp16 with a per-device batch size of 2 and gradient accumulation of 16, using AdamW with a cosine learning-rate schedule and 5% warmup, for 2 epochs on the official parallel data.

4.3 X-ALMA SFT (QLoRA)

The X-ALMA-13B backbone is loaded in 4-bit NF4. We attach a fresh LoRA ($r=16$, $\alpha=32$) on all linear projections, with the loss masked to target tokens only, and optimize with paged AdamW (8-bit) and gradient checkpointing.

4.4 CPO preference fine-tuning

On top of the X-ALMA SFT adapter we run Contrastive Preference Optimization (CPO) (Xu et al., 2024). For each training source we sample $K=4$ candidates from the SFT model, score the pool augmented with the gold reference using CometKiwi (Rei et al., 2022b), and form preference triples by pairing the highest-scoring candidate (chosen) against the lowest-scoring candidate (rejected). CPO modestly improves BLEU and

COMET over X-ALMA-FT in both directions, at a small cost in TER for en→ar.

4.5 ALLaM fine-tuning

ALLaM-7B-Instruct is fine-tuned per direction with QLoRA (4-bit NF4 quantization, rank $r=64$, $\alpha=128$, all linear projections as targets) on the official parallel data, with a maximum sequence length of 1024, a per-device batch size of 8, and gradient accumulation of 4. FlashAttention-2 (Dao, 2023) is enabled where available.

4.6 Decoding and ensembling

Single-system inference uses beam search (beam size 4) for the sequence-to-sequence models and greedy decoding for the LLM systems. The best-1 outputs feed the system-combination MBR of Section 3.1; the sampling-MBR procedure is as described in Section 3.3.

5 Experimental Setup

Development-set scores are computed on the official development set ($n=1000$ per direction). BLEU, chrF, and TER are computed with sacreBLEU 2.6.0 (Post, 2018; Papineni et al., 2002; Popović, 2015; Snover et al., 2006) using default tokenization (signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0`); COMET is the `wmt22-comet-da` model (Rei et al., 2022a, 2020), and CometKiwi (`wmt22-cometkiwi-da`) (Rei et al., 2022b) is used for QE pre-pruning. All training and inference were performed on $3 \times$ NVIDIA RTX 3090 (24 GB each); the full pipeline takes roughly two days end-to-end.

The challenge test set is blind (1,500 segments per direction, no released references) and is scored by the organizers using BLEU, chrF, TER,

Table 1: Sub-Task 2A (Arabic→English) development-set results. P: primary; C1/C2: contrastive. Best per column in **bold**.

System	BLEU	chrF	TER	COMET
opus-mt	26.49	54.08	62.36	0.8222
X-ALMA pretrain	24.67	52.30	64.09	0.8380
NLLB-3.3B FT	31.51	57.86	58.64	0.8447
X-ALMA-FT	31.76	58.23	59.18	0.8449
X-ALMA-CPO	32.23	58.81	58.54	0.8468
ALLaM-FT (C1)	33.06	59.18	58.56	0.8449
Sampling-MBR (C2)	30.06	57.26	59.33	0.8524
MBR sys-combo (P)	32.97	59.10	57.32	0.8531

Table 2: Sub-Task 1A (English→Arabic) development-set results. P: primary; C1/C2: contrastive. Best per column in **bold**.

System	BLEU	chrF	TER	COMET
X-ALMA pretrain	14.94	49.07	78.28	0.8624
X-ALMA-FT	17.38	49.35	72.84	0.8610
X-ALMA-CPO	17.77	50.44	74.35	0.8630
opus-mt	18.84	52.87	72.15	0.8678
NLLB-3.3B FT	21.19	54.58	68.28	0.8789
ALLaM-FT (C1)	21.74	54.40	69.10	0.8750
Sampling-MBR (C2)	18.72	52.74	71.26	0.8862
MBR sys-combo (P)	21.48	54.90	68.97	0.8865

COMET, and BERTScore. We submitted the primary and two contrastive systems for both directions; the official results are reported in Section 7.

6 Development-Set Results

Tables 1 and 2 report development-set results. The system-combination MBR primary is the strongest system overall: it attains the best COMET in both directions (0.8531 for ar→en and 0.8865 for en→ar), although its COMET margin over sampling-MBR is not statistically significant (Section 8). It also obtains the best TER for ar→en and the best chrF for en→ar, and trails the top BLEU system (ALLaM-FT) by only 0.09 and 0.26 BLEU respectively. Among the single systems, ALLaM-FT gives the best BLEU and sampling-MBR the best COMET in both directions.

7 Official Test-Set Results

Tables 3 and 4 report the organizers’ official scores on the blind challenge test set, alongside the strongest competing team for context. Note that COMET is reported on the 0–100 scale here, matching the official scoreboard, whereas the development-set tables use the 0–1 scale; primary and contrastive runs are ranked on separate leader-

Table 3: Official test-set results, Sub-Task 2A (Arabic→English). Our primary ranks first on the primary leaderboard; NLP-IIT Patna is the runner-up. Contrastive-1 is ALLaM-FT, Contrastive-2 is sampling-MBR. Best per column among the shown rows in **bold**. COMET and BERTScore on the 0–100 scale. The official 2A contrastive scoreboard lists BERTScore identical to COMET for every row, which we treat as unreported (–).

System	BLEU	chrF	TER	COMET	BERT
Ours (P)	33.83	59.72	57.00	82.49	83.30
Ours (C1)	34.02	59.98	58.51	82.61	–
Ours (C2)	31.26	58.19	58.82	82.71	–
NLP-IIT Patna (P)	33.67	59.67	57.30	82.01	83.11

Table 4: Official test-set results, Sub-Task 1A (English→Arabic). Our primary ranks second by BLEU but attains the highest COMET on the primary leaderboard. Contrastive-1 is ALLaM-FT, Contrastive-2 is sampling-MBR. Best per column among the shown rows in **bold**. COMET and mBERTScore on the 0–100 scale.

System	BLEU	chrF	TER	COMET	mBERT
NLP-IIT Patna (P)	22.45	55.17	67.26	84.74	85.67
Ours (P)	21.81	54.99	69.29	85.23	85.40
Ours (C1)	22.48	54.58	68.21	84.26	85.31
Ours (C2)	19.17	52.59	71.17	84.93	84.87

boards.

For Arabic→English (Table 3), our primary system-combination ranks **first** on the primary leaderboard, with the best BLEU, chrF, and TER among all teams. Both contrastive runs edge the primary on COMET (82.61 and 82.71 versus 82.49) while trailing it on TER. For English→Arabic (Table 4), our primary ranks second by BLEU but attains the highest COMET of any system on the primary leaderboard (85.23 versus 84.74 for the top BLEU system), and our contrastive-1 (ALLaM-FT) tops the contrastive leaderboard on BLEU. That the contrastive systems overtake the primary on COMET for ar→en is consistent with the development set, where the primary’s COMET margin over sampling-MBR was already within bootstrap noise (Section 8).

8 Discussion

Selection metric versus evaluation metric. COMET-22 serves both as the MBR utility and as one of the reported evaluation metrics, so part of the primary’s COMET advantage is expected: the system is selected to maximize expected COMET, and a COMET gain over the constituents is partly

a consequence of that objective rather than independent evidence of quality. We therefore read the COMET column with this caveat, and rely on metrics that play no role in selection to judge whether the improvement is real. Two such signals are available. First, the lexical metrics (BLEU, chrF, and TER) are not used anywhere in the pipeline; on the official ar→en test set the primary is best among all teams on all three (Table 3), and on en→ar it is within 0.6 BLEU of the top system (Table 4). Second, the organizers’ BERTScore is an embedding-based metric independent of COMET; on ar→en the primary is again the best primary submission (83.30), and on en→ar it is second (85.40 versus 85.67). The primary’s advantage is thus corroborated by metrics it was not optimized for, most clearly in the ar→en direction. We did not conduct a human assessment; the organizers’ native-speaker human evaluation, where released, is the appropriate independent check.

Why combination over a single best model.

The single fine-tuned systems disagree about which metric they optimize: the Arabic-native LLM is strongest on the surface-overlap metric (BLEU), while sampling-MBR with a neural utility is strongest on the embedding-based metric (COMET). This is the classic tension between lexical and semantic evaluation. Rather than simply hedging it, system-combination MBR exploits it: by letting a quality estimator arbitrate per sentence over a pool of diverse hypotheses, the primary matches or exceeds every single constituent on COMET in both directions while staying within a few tenths of a point of the best BLEU and chrF. The diversity of the pool is therefore a resource the combination converts into a net quality gain.

Which members of the pool matter. Table 5 reports a leave-one-out ablation on the development set: the primary is re-decoded with each system removed from the pool, all else equal. Two systems carry the combination. Removing ALLaM-FT costs the most COMET in both directions (−0.0019 and −0.0034) and 1.0 BLEU for en→ar; removing NLLB-3.3B FT costs a similar 1.1 BLEU and 0.0030 COMET for en→ar. The remaining members contribute little once these two are present: dropping X-ALMA-CPO leaves COMET unchanged and slightly *raises* BLEU for en→ar (+0.17), dropping X-ALMA-FT is neutral, and dropping OPUS-MT is marginal for en→ar and slightly beneficial for ar→en (+0.12 BLEU). The

Table 5: Leave-one-out ablation of the MBR candidate pool on the development set. Each row removes one system and re-runs MBR; the first row is the submitted primary. COMET on the 0–1 scale; chrF and TER follow the same pattern.

Removed	ar→en		en→ar	
	BLEU	COMET	BLEU	COMET
none (full)	32.97	0.8531	21.48	0.8865
NLLB-3.3B FT	32.91	0.8519	20.42	0.8835
X-ALMA-FT	32.91	0.8527	21.52	0.8863
X-ALMA-CPO	32.61	0.8527	21.65	0.8865
ALLaM-FT	32.56	0.8512	20.45	0.8831
opus-mt	33.09	0.8532	21.10	0.8852

useful diversity is therefore not “more translation LLM variants” but the pairing of an Arabic-native decoder-only LLM with a multilingual encoder-decoder, whose outputs are complementary enough for COMET-MBR to arbitrate between them. Consistently, enlarging an earlier three-system pool (X-ALMA-FT, NLLB-3.3B FT, OPUS-MT) to five improved every metric in both directions (+1.19 and +0.53 BLEU), a gain attributable mainly to ALLaM-FT.

Are the differences significant? Table 6 reports paired bootstrap resampling (Koehn, 2004) on the development set (1,000 resamples, sentence-level COMET-22 for the COMET test), comparing the primary against its strongest constituents and against Contrastive 2. The official test set does not permit this test because its references are not released, so we treat the development set as the closest available proxy. The picture is consistent across directions. Against sampling-MBR, the primary’s COMET advantage is *not* significant ($p=0.10$ and 0.24): on the metric it optimizes, the two are statistically tied, while the primary is far better on BLEU, chrF and TER ($p\leq 0.001$). Against ALLaM-FT the pattern inverts: ALLaM’s BLEU edge is not significant, whereas the primary’s COMET gain is ($p=0.001$). The primary is thus the only system that is never significantly worse than another on any metric, with a single exception: NLLB-3.3B FT has significantly lower TER for en→ar ($p=0.016$). We read this as the practical case for combination in a shared-task setting where the deciding metric is unknown. We do not claim significance over the strongest competing *team*: on the blind test set the primary leads the ar→en runner-up by 0.16 BLEU and 0.49 COMET, margins we cannot test.

Table 6: Paired bootstrap on the development set. Δ is primary minus system; positive favors the primary for BLEU, chrF and COMET, and negative favors it for TER. †: not significant at $p<0.05$; all other entries have $p<0.05$ (most $p\leq 0.006$).

Dir.	Primary vs.	Δ BLEU	Δ chrF	Δ TER	Δ COMET
ar→en	ALLaM-FT	-0.09†	-0.08†	-1.24	+0.0082
	X-ALMA-CPO	+0.74	+0.29	-1.23	+0.0063
	NLLB-3.3B FT	+1.46	+1.24	-1.32	+0.0084
	Sampling-MBR	+2.91	+1.84	-2.01	+0.0007†
en→ar	ALLaM-FT	-0.26†	+0.49	-0.14†	+0.0115
	X-ALMA-CPO	+3.72	+4.46	-5.38	+0.0235
	NLLB-3.3B FT	+0.30†	+0.31	+0.69	+0.0076
	Sampling-MBR	+2.76	+2.15	-2.29	+0.0002†

The value of an Arabic-native LLM. ALLaM-7B was pretrained with a deliberately Arabic-centric mixture (Bari et al., 2025), and a light QLoRA fine-tune on the official data is enough to make it the highest-BLEU single system in both directions and, by the leave-one-out ablation, the single most valuable member of the pool. This suggests that, in a low-resource setting, a model whose pretraining already covers the target language can be more sample-efficient to adapt than a more general multilingual model.

Quality estimation as a decoding signal. Both the sampling-MBR system and the primary ensemble rely on neural quality estimators (CometKiwi for pre-pruning, COMET-22 for selection) to pick outputs. Using QE to prune the candidate pool before the more expensive pairwise MBR scoring keeps the procedure tractable at $K=32$.

9 Conclusion

We presented OpenBracket’s submission to the WMT 2026 Low-Resource Arabic–Asian translation task for the English-paired directions. Our primary system combines diverse, individually strong models via COMET-MBR; on the development set it attains the best COMET in both directions and is never significantly worse than any constituent on any metric save one, and on the official blind test set it ranks first in Arabic→English (best BLEU, chrF, and TER) and obtains the highest COMET on the English→Arabic primary leaderboard. Among the single systems, an Arabic-native LLM gives the best BLEU and sampling-MBR over a multilingual model gives the best COMET; the combination converts this diversity into a net gain, and a leave-one-out ablation shows that the Arabic-native LLM and the multilingual encoder-decoder are the pool

members that matter. We release our code and decoding scripts to support reproduction.¹

Limitations

We report both development-set and official test-set results. As the official scores show, system rankings shift somewhat between dev and test, particularly between lexical and neural metrics, so the relative ordering of our systems should be read with that metric dependence in mind. The official test-set scores are single-run point estimates on a blind set whose references are not released, so significance testing between the leading *teams* is not possible; the margins between the top primary submissions are small (for example 0.16 BLEU for ar→en) and should not be over-interpreted. Our significance tests and pool ablation (Tables 5 and 6) are on the development set only, so they characterize our own systems rather than the official ranking. COMET-22 is both our selection objective and a reported metric, a circularity we discuss in Section 8. MBR and sampling-MBR are computationally heavier than single-pass decoding, which may limit applicability under tight latency budgets. We evaluate only the two English-paired directions (1A and 2A) of the task and do not cover the Hindi, Bangla, Indonesian, or Urdu pairs. Finally, our selection metrics (COMET-22 and CometKiwi) were themselves trained on data that may not fully reflect the dialects and domains of this low-resource benchmark.

References

- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Al-Rashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Al-rubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2025. *ALLaM: Large language models for arabic and english*. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. ArXiv:2407.15390.
- Tri Dao. 2023. *FlashAttention-2: Faster attention with better parallelism and work partitioning*. *Preprint*, arXiv:2307.08691.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. *QLoRA: Efficient fine-tuning of quantized LLMs*. In *Advances in*

¹<https://github.com/AbrarAlotaibi/AsianMT>

- Neural Information Processing Systems (NeurIPS)*. ArXiv:2305.14314.
- Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4506–4520.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics (TACL)*, 10:811–825.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.
- 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 (ICLR)*. ArXiv:2106.09685.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.
- 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.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and 1 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- 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 (ACL)*, 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 (WMT)*, pages 392–395.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 186–191.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, pages 479–480.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 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.
- Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. [X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale](#). In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. ArXiv:2410.03115.
- 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](#). In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. ArXiv:2401.08417.