HexaBee-4B: Six-Language Translation with Multilingual Answering and Bidirectional Supervision

Iman Barati[1]* **Mohammadmostafa Rostamkhani**[1] **Mohammad Badzohreh**[2]

[1]Iran University of Science and Technology, Tehran, Iran

[2]Tarbiat Modares University, Tehran, Iran

iman_barati@comp.iust.ac.ir mo_rostamkhani97@comp.iust.ac.ir

mohammad.badzohreh@modares.ac.ir

Abstract

Multilingual translation requires a model to preserve meaning while producing text in a requested language. We present HEXABEE-4B, a translator adapted from Gemma 3 4B IT through three stages: Wikipedia continued pretraining, joint multilingual answering and answer-text translation, and bidirectional translation specialization. The intermediate stage connects two tasks through shared multilingual answers: answering a question in the requested language and translating an answer into that language. It comprises 19.46 million answering records, including 5.76 million cross-language question–answer pairs, and 10.70 million answer-text translations. The final stage uses 658,274 forward and reverse synthetic translation examples. On 36,572 synthetic test examples spanning all 30 directions among Arabic, Chinese, English, Hebrew, Persian, and Spanish, HEXABEE-4B achieves direction-macro chrF++ of 73.75, compared with 57.11 for Gemma 3 4B IT and 66.83 for Gemma 3 12B IT. BLEU, TER, COMET, and a matched model-judged subset support the same overall ranking. These results demonstrate improved translation across the six-language set, with chrF++ gains in every evaluated direction over both baselines.

1 Introduction

A multilingual language model can encounter the same information in several languages, yet translation requires more than recognizing that information: its response must preserve the source meaning and follow a specified target language. Multilingual neural translation establishes target-language conditioning as a central design choice (Johnson et al., 2017). More recent work shows that continued training and translation-oriented instruction tuning can adapt general language models into stronger translators (Xu et al., 2024; Li et al., 2024). For a compact instruction-tuned model, the organization of this supervision is therefore an important part of the translation system.

Multilingual instruction tuning supplies a broader source of supervision. Multitask training and multilingual instruction collections improve instruction following across languages (Muennighoff et al., 2023; Singh et al., 2024; Üstün et al., 2024); even limited multilingual instruction data can support cross-language generalization (Shaham et al., 2024). However, multilingual data need not require communication *between* languages. A question and its answer may always use the same language. Explicitly separating their language assignments creates a different task: understand a question in one language and answer it in another. Cross-language instruction tuning and translation-following demonstrations provide precedents for this design (Lin et al., 2025; Ranaldi et al., 2024; Zhu et al., 2024).

We investigate how to organize such answering supervision within a translation-focused adaptation recipe (Figure 1). Consider an English question whose requested answer is Spanish. The model must answer the question in Spanish, whereas an English-to-Spanish translation task must preserve the content of its English input. Our construction connects these tasks through multilingual answer texts: a Spanish answer can supervise both an English question and an English version of that answer. Figure 2 illustrates this shared-target relation. The system instruction specifies the desired response language and, where relevant, response style. Its own wording can vary across languages, but this variation is a supporting data-construction choice. Translation remains the primary task evaluated here.

This separation of language roles is also motivated by multilingual question-answering research. TyDi QA and MKQA broaden the linguistic cov-

*Corresponding author.

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 46–70
November 9, 2026 ©2026 Association for Computational Linguistics

erage of QA evaluation (Clark et al., 2020; Longpre et al., 2021). XHotpotQA makes language assignments to questions and evidence roles explicit when studying cross-lingual knowledge composition (Barati et al., 2026). Our work applies the related principle of separating language roles to training question–answer generation and answer-text translation. The resulting supervision uses a Wikipedia-derived multilingual inventory and does not require retrieving evidence at inference time.

We introduce HEXABEE-4B, built from **Gemma 3 4B IT** (Gemma Team, 2025).[1] Its three stages provide encyclopedic language modeling, multilingual response supervision, and translation specialization. The second stage combines 19,458,332 answering records with 10,695,072 answer-text translations; 5,759,806 answering records cross question and response languages. The third stage uses 658,274 bidirectional synthetic examples. We evaluate the complete model across six languages using five automatic metrics and paired model-based judgments. Our contributions are the three-stage translation recipe, an explicit account of how answering and translation share supervision, and a direction-complete evaluation against public Gemma 3 4B IT and 12B IT. The comparison measures the complete recipe; it does not isolate the effect of an individual training component.

2 Related Work

Multilingual translation and adaptation. Multilingual MT shares parameters across translation directions and uses explicit target-language signals (Johnson et al., 2017). ALMA studies translation-oriented continued training of large language models (Xu et al., 2024), while Li et al. (2024) examine multilingual finetuning with translation instructions. Our staged design combines this translation focus with an intermediate answering task. Its initial Wikipedia adaptation follows the broader motivation of continued pretraining on relevant text (Gururangan et al., 2020). Training on teacher-generated translations is related to sequence-level knowledge distillation (Kim and Rush, 2016).

Multilingual instruction supervision. Multitask finetuning enables cross-lingual generalization (Muennighoff et al., 2023), and Aya provides multilingual instruction data and models (Singh et al., 2024; Üstün et al., 2024). CrossIn combines cross-language instruction–response pairs with instruction translation (Lin et al., 2025). CrossAlpaca uses instruction-following and translation-following demonstrations (Ranaldi et al., 2024); Question Translation Training aligns reasoning questions across languages (Zhu et al., 2024). These studies establish relevant precedents for coupling instruction following and translation. Our contribution is the particular combination of system-directed answering, translations of answer texts, and subsequent bidirectional specialization on Gemma 3 4B IT, together with its measured corpus composition and translation results.

Language roles in question answering. TyDi QA collects information-seeking questions directly in diverse languages (Clark et al., 2020), while MKQA supports comparison across languages through aligned questions and language-independent answer representations (Longpre et al., 2021). XHotpotQA studies mixed-language evidence in multi-hop QA by assigning languages to distinct roles in an evidence-dependency graph (Barati et al., 2026). This provides a conceptual motivation for making language boundaries explicit rather than treating multilinguality only as a collection of monolingual examples. Here, the relevant roles are the question, requested answer, and instruction wording; we use their combinations as generation supervision for a translator. Our experiments concern translation quality, whereas those benchmarks evaluate answering or knowledge composition.

3 Method

3.1 Answering and translation as connected tasks

Let $\mathcal{L} = \{\mathrm{ar}, \mathrm{zh}, \mathrm{en}, \mathrm{he}, \mathrm{fa}, \mathrm{es}\}$ be the six translation languages. We use $s, t \in \mathcal{L}$ for input and response languages. For a content item c, let $q_s(c)$ denote its question in language s, and let $a_t^v(c)$ denote an answer in language t with response variant v (extent or style). Let S denote the *system instruction text*. Specifically, $S_{t,v}^{\mathrm{QA}}$ requests an answer in language t with variant v, and $S_{s,t}^{\mathrm{MT}}$ requests translation from s into t.

[1] Resources: Code; Final model.

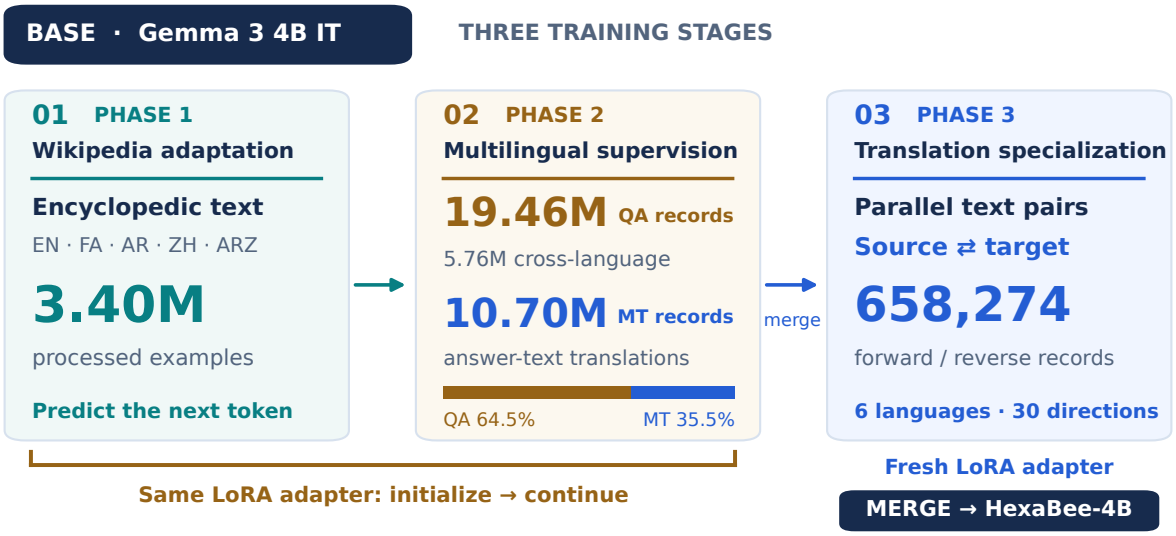


Figure 1: The complete adaptation recipe. Phase 1 applies next-token language modeling to Wikipedia text. Phase 2 continues the same adapter with multilingual answering and answer-text translation. After merging, phase 3 trains a fresh adapter on bidirectional translations and produces the final merged HexaBee-4B model. The base is Gemma 3 4B IT. Phase 1 counts processed training examples; phases 2–3 count SFT records. Examples derived from one source can occur in multiple tasks.

The two supervision relations are

$$\text{QA}: \quad (S^{\text{QA}}_{t,v}, q_s(c)) \mapsto a^v_t(c), \tag{1}$$

$$\text{MT}: \quad (S^{\text{MT}}_{s,t}, a_s(c)) \mapsto a_t(c), \tag{2}$$

where $a_s(c)$ and $a_t(c)$ are the answer variants selected for translation. QA includes same-language and cross-language pairs; MT uses different input and output languages. A shared answer can therefore appear as the target of both tasks. Its input is a question in QA and an answer text in MT. This relation connects response generation to semantic transfer without adding an intermediate generation step.

Instruction pools provide alternative phrasings for response language, extent, style, and translation direction.[2] Cross-language QA explicitly requests the answer language; some same-language examples request only answer extent. A directive's wording language and its requested response language are separate choices. We preserve this distinction when constructing the data, rather than assuming that a prompt written in one language always requests that language. Pool composition and sampling are documented in Appendix D.

3.2 A shared response objective

For either task, let x be the input text and $y = (y_1, \ldots, y_m)$ the target response of m tokens. With

[2]Resources: System prompts.

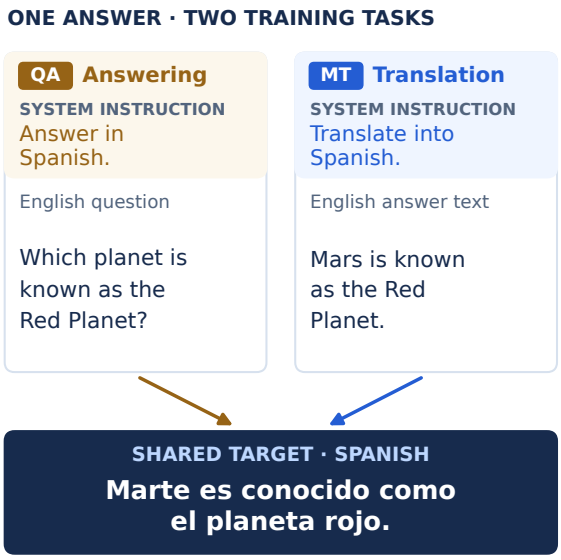


Figure 2: Illustrative English-to-Spanish supervision. The tasks have different inputs and instructions but can share the same target answer. This constructed example explains the method; it is not a quoted training record.

model parameters θ, the response-token negative log-likelihood is

$$\ell_\theta(S, x, y) = -\sum_{k=1}^{m} \log P_\theta(y_k \mid S, x, y_{<k}), \tag{3}$$

where $y_{<k}$ is the response prefix preceding token k. The instruction and input are condition-

ing text; the response supplies supervision. QA and translation examples use this same objective with their respective inputs and targets. Let $\mathcal{D}_{\mathrm{QA}}$ and $\mathcal{D}_{\mathrm{MT}}$ denote the retained QA and translation records. The phase-2 dataset is the multiset union $\mathcal{D}_2 = \mathcal{D}_{\mathrm{QA}} \uplus \mathcal{D}_{\mathrm{MT}}$, so repeated or reused examples retain their contribution to the training mixture. We do not introduce a separate alignment loss between the tasks.

4 Three-Stage Training Data and Adaptation

4.1 Phase 1: Wikipedia continued pretraining

The first stage adapts Gemma 3 4B IT to encyclopedic text before introducing the paired generation tasks. It uses Wikipedia text in English, Persian, Arabic, Chinese, and Egyptian Arabic. These are the languages registered for this stage; the later six-language translation set additionally includes Hebrew and Spanish. The corpus is monolingual text, so this phase provides language-modeling supervision rather than question–answer or translation pairs.

For a tokenized sequence $w = (w_1, \ldots, w_n)$, its autoregressive objective is

$$\ell_\theta^{\mathrm{PT}}(w) = -\sum_{k=2}^{n} \log P_\theta(w_k \mid w_{<k}), \qquad (4)$$

where n is sequence length and $w_{<k}$ contains preceding tokens. In contrast to Equation 3, supervision is applied throughout the text sequence, without a separate requested response. This stage aims to adapt vocabulary use and encyclopedic text modeling to the material used in the next stage, consistent with the motivation of continued pretraining (Gururangan et al., 2020).

A LoRA adapter (Hu et al., 2022) is trained for one epoch on 3,399,664 processed training examples. The available raw source files contain 11,395,061 text rows before preprocessing; these counts have different units. We retain the trained adapter for phase 2.[3] Appendix A gives source-language counts and the repeated-source composition of the original mixture.

4.2 Phase 2: multilingual answering and answer translation

The second stage constructs multilingual question and answer variants from 2,025,730 Wikipedia-derived source rows. The source inventory begins with Persian content and is expanded into the other languages using Gemma 3 27B IT (Gemma Team, 2025). Figure 3 traces this inventory into the two supervision branches. The resulting multilingual inventory, rather than a single language, defines the supervision in Equations 1–2. Construction provenance and translation paths are given in Appendix B.

Phase-2 supervision	Records
Same-language answering	13,698,526
Cross-language answering	5,759,806
Answer-text translation	10,695,072
Total	30,153,404

Table 1: Retained phase-2 supervision. Same/cross refers to question and answer language. Response-style and direction breakdowns are given in the appendix.

For answering, the input is a question and the target is an answer selected in the requested language. The variants include complete and moderate answers, as well as formal and informal responses. For answer-text translation, both sides are answer variants associated with the same content. All constructed answer-translation groups participate in this phase alongside QA; translation is therefore a substantial part of the intermediate training, not only the final specialization.

The retained mixture contains 30,153,404 records[4]: 19,458,332 answering records and 10,695,072 answer-text translations (Table 1). Cross-language answering accounts for 5,759,806 records, or 29.6% of QA supervision. Answering and translation account for 64.5% and 35.5% of the full mixture, respectively. These are corpus proportions, not additional loss weights. The same source item can contribute to several tasks or language combinations.

The QA distribution is asymmetric: the complete-answer family covers 27 cross-language directions, while the combined translation data cover all 30. Retained data also contain empty fields and a small set of conflicting language directives. We document their counts and affected directions in Appendix C. The intended task is language-controlled answering; the recorded corpus is an imperfect realization of that design.

Phase 2 continues the phase-1 adapter for one epoch on the combined answering/translation mix-

[3] Resources: Wikipedia corpus; Phase-1 adapter.

[4] Resources: Phase-2 training data.

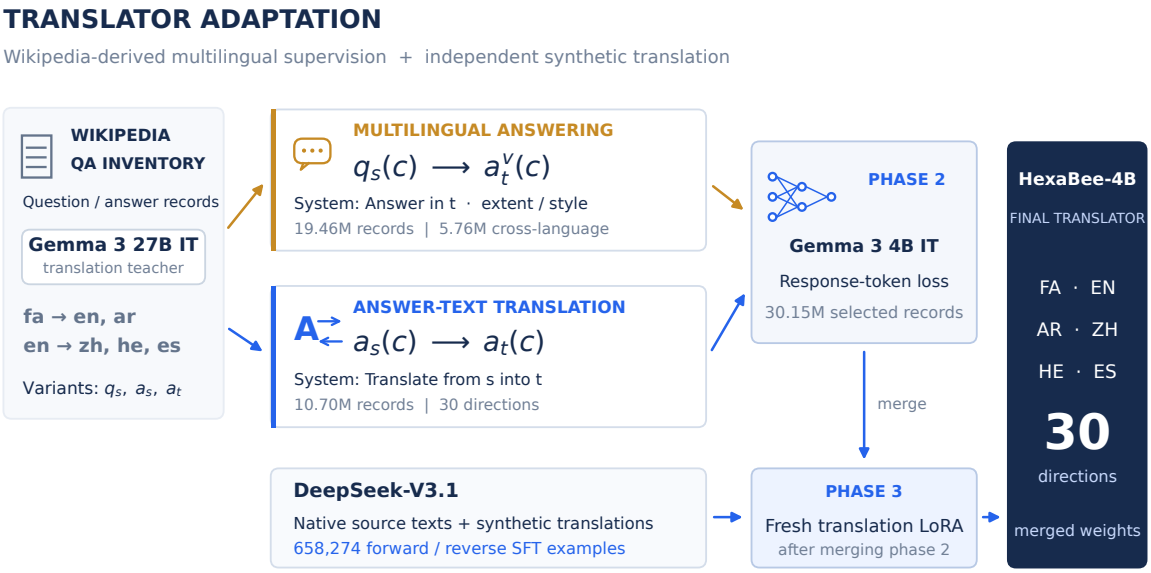


Figure 3: Data and supervision flow in phases 2 and 3. Gemma 3 27B IT expands the Wikipedia-derived inventory into multilingual question and answer variants. Answering and answer-text translation supply complementary phase-2 inputs, with system instructions specifying the requested response. A separate DeepSeek-generated corpus supplies the final bidirectional translation stage. The full checkpoint sequence, including phase 1, is shown in Figure 1.

ture. The resulting adapter is then merged with Gemma 3 4B IT to obtain the multilingual checkpoint used by the final stage.[5] This continuity allows both phases to contribute to the checkpoint before translation specialization.

4.3 Phase 3: bidirectional translation specialization

The third stage uses a separately generated translation corpus. DeepSeek-V3.1 source-generation metadata (DeepSeek-AI, 2025, 2024) identifies the teacher used to produce texts across topics, genres, registers, and complexity levels in the six languages. Each source text is paired with a synthetic translation into a different language. Native source generation broadens the content beyond the Wikipedia-derived answering inventory; its diversity is a construction goal, not evidence of coverage of every deployment domain.

The corpus contains 365,718 aligned pairs, with 60,953 source texts per language. Its stored split contains 329,146 training pairs and 36,572 test pairs. Filtering leaves 329,137 training pairs for bidirectional supervision. Each retained pair supplies two examples: translating the source into its paired target and translating the paired target back into the original source language. Thus the final SFT dataset contains $2 \times 329{,}137 = 658{,}274$ records.[6] Both orientations share an underlying pair. The reverse example reuses the existing texts; it does not require another teacher call.

The use of decoded teacher translations is related to sequence-level distillation (Kim and Rush, 2016). Here, both source generation and target translation are synthetic, so evaluation on the held-out portion measures this particular distribution. The exact split and filtering details are retained in Appendix B.

A fresh LoRA adapter is trained on the merged phase-2 checkpoint using the response objective in Equation 3. After one epoch, it is merged to produce HEXABEE-4B.[7] All three stages use LLaMA-Factory (Zheng et al., 2024); adapter parameterization, optimization settings, and serialization are in Appendix E. The complete sequence is therefore continued pretraining, joint answering/translation SFT, and translation-only SFT, with two adapter merges.

[5] Resources: Phase-2 adapter; Merged multilingual checkpoint.

[6] Resources: Bidirectional translation corpus.

[7] Resources: Phase-3 adapter.

5 Experimental Setup

5.1 Evaluation data and comparison systems

We evaluate stored outputs for all 36,572 held-out synthetic examples, covering the 30 directed language pairs. Each direction contains 1,173–1,276 examples. The comparison includes HEXABEE-4B, public Gemma 3 4B IT, and public Gemma 3 12B IT (Gemma Team, 2025); all three use the same source texts and references. A matched subset of 300 examples, ten per direction, is additionally scored by a model judge.

The archived inference templates differ between HEXABEE-4B and the public baselines. Consequently, results compare the recorded systems, including their prompts, rather than isolating adaptation under an identical prompt. We score outputs as stored, including empty responses and unsolicited text. Inference settings and failure counts are reported in Appendix B.4. No human-authored translation benchmark is evaluated in this study; FLORES provides a complementary evaluation setting for future validation (Goyal et al., 2022).

5.2 Automatic metrics

We report BLEU (Papineni et al., 2002), chrF (Popović, 2015), chrF++ (Popović, 2017), TER (Snover et al., 2006), and COMET (Rei et al., 2020). SacreBLEU provides reproducible lexical-metric settings (Post, 2018); COMET uses the fixed `wmt22-comet-da` model (Unbabel, 2022).[8] TER is lower-is-better; the other reported metrics are higher-is-better. COMET values are multiplied by 100 for display.

Let $\Omega = \{(s,t) \in \mathcal{L}^2 : s \neq t\}$ be the set of 30 directions and M_d the score of a metric computed on direction d. Our primary summary is

$$M_{\text{macro}} = \frac{1}{|\Omega|} \sum_{d \in \Omega} M_d. \tag{5}$$

This gives every direction equal weight. A target-language summary averages its five incoming directions. Pooled corpus scores are different aggregations and are reported separately in Appendix F, together with metric settings, encoder truncation, and every directional result.

5.3 Model judging and uncertainty

Model-based translation evaluation can complement reference-overlap metrics (Kocmi and Federmann, 2023b,a). We use the archived DeepSeek judge in two modes: reference-assisted, with the synthetic reference, and reference-free, with source and candidate only. Both modes produce a holistic score from 0 to 100 and eight rubric dimensions covering accuracy, terminology, fluency, cohesion, context, language adherence, style, and formatting. These are model judgments, not calibrated accuracy percentages or human annotations.

We compare systems on the same 300 examples and estimate 95% intervals for score differences using paired bootstrap resampling within each translation direction (Koehn, 2004). This preserves the matched comparison and gives each direction the same weight. The intervals capture item-sampling uncertainty on this subset; they do not quantify variability across training runs or alternative judges. Appendix B.5 gives the procedure and rubric details.

6 Results and Analysis

6.1 Overall translation quality

HEXABEE-4B obtains the best overall score on all five automatic metrics (Table 2). Direction-macro chrF++ reaches 73.75, compared with 57.11 for Gemma 3 4B IT and 66.83 for Gemma 3 12B IT. This is a gain of 16.64 points over the same-size public base and 6.92 points over the larger comparator. BLEU increases from 40.70 to 62.03 relative to the 4B baseline; TER decreases from 51.40 to 30.05, and COMET increases from 85.88 to 89.86. The agreement between lexical and learned metrics supports improved translation on the evaluated synthetic distribution.

6.2 Language and direction differences

The chrF++ gain is positive in every direction relative to both baselines. Figure 4 shows improvements for each target language, while Appendix F reports all source–target comparisons. Relative to the 4B baseline, the largest directional gain is Chinese-to-Persian (+27.09), and the smallest is English-to-Spanish (+6.64). Thus improvement extends beyond one target language or one script pairing.

Absolute chrF++ remains lower for Chinese-target translations than for the other target groups. This should not be read as a calibrated ranking of language difficulty: script, segmentation, reference phrasing, and the metric's character/word

[8] Resources: Stored model outputs; Evaluation tables and figures.

System	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.
4B IT	40.70	58.91	57.11	51.40	85.88	80.32	81.43
12B IT	50.54	68.70	66.83	38.89	88.20	91.73	91.97
HEXABEE-4B	**62.03**	**75.08**	**73.75**	**30.05**	**89.86**	**92.92**	**94.10**

Table 2: Full-test direction-macro automatic metrics ($N = 36{,}572$) and holistic judge means on 300 matched items. COMET is scaled by 100; TER is lower-is-better. Bold marks the best displayed score and underlining the second distinct score. Each system uses its recorded inference prompt.

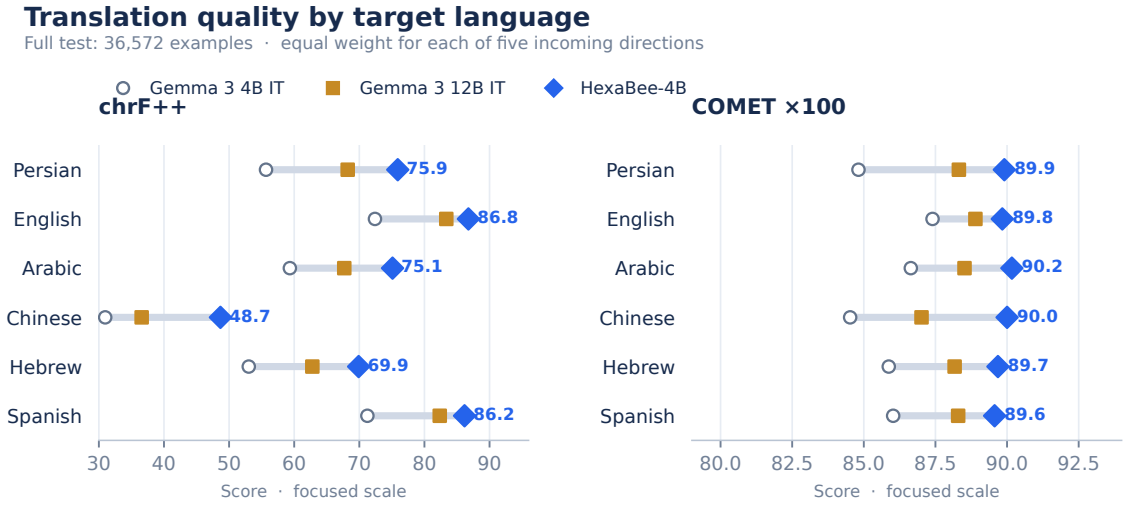


Figure 4: Translation quality by target language. Each point averages five incoming directions on the full test partition. HEXABEE-4B improves both direction-macro chrF++ and COMET for each target language. The focused axis ranges are shown explicitly; marker shapes distinguish systems as well as colors.

components affect cross-language score comparisons. The metric agreement is also not universal at the direction level. For Chinese-to-Hebrew, HEXABEE-4B has TER 90.09, worse than 88.40 for 4B IT and 78.13 for 12B IT, despite higher BLEU and chrF++. TER improves in the other 29 directions. This exception motivates inspecting multiple metrics rather than inferring uniform improvements from a single aggregate.

6.3 Judged quality

Both judge modes follow the automatic-metric ranking. Reference-assisted means are 92.92 for HEXABEE-4B, 80.32 for 4B IT, and 91.73 for 12B IT; reference-free means are 94.10, 81.43, and 91.97. Table 3 reports paired differences and uncertainty. The comparison with the larger model is considerably closer than the comparison with the 4B base.

Against 4B IT, HEXABEE-4B wins, ties, or loses on 214, 54, and 32 reference-assisted examples, respectively. Across rubric dimensions, its terminology mean increases from 75.60 to 90.42, and target-language adherence from 82.88 to 96.15. Terminology nevertheless remains its lowest reference-assisted dimension. These judgments suggest a useful focus for further bilingual error analysis; they do not establish the frequency of specific error categories. Complete rubric results by direction and target language are included in the appendix.

Baseline	Judge view	Gain	95% CI
4B IT	Reference	+12.60	[9.96,15.31]
4B IT	No reference	+12.67	[10.13,15.34]
12B IT	Reference	+1.19	[0.34,2.08]
12B IT	No reference	+2.13	[1.29,3.13]

Table 3: Paired holistic-score differences, HEXABEE-4B minus baseline, on the matched 300-item subset.

7 Discussion

The method combines three forms of supervision with different roles. Continued pretraining adapts language modeling to encyclopedic text. Multilingual answering requires selecting the information requested by a question and expressing it in a specified language; answer-text translation exposes the

model to semantic counterparts in that same response space. Final bidirectional SFT then concentrates training on the translation relation evaluated at test time. Shared answer targets provide a concrete connection between the intermediate tasks, while the final stage uses a separate source-generation process.

The results show that this complete recipe produces a stronger translator on the recorded test set. They do not establish which component causes the gains. A controlled attribution study would keep the Wikipedia-adapted checkpoint, total training budget, final translation corpus, and inference prompts fixed while varying the intermediate answering/translation mixture. Evaluating the intermediate checkpoint would further distinguish improvements before and after specialization. These comparisons are not available here. The present contribution is a fully specified system and its translation evaluation, with corpus counts that make those future comparisons possible.

8 Conclusion

We presented HEXABEE-4B, a six-language translator adapted from Gemma 3 4B IT through Wikipedia continued pretraining, multilingual answering with answer-text translation, and bidirectional translation specialization. The central data design links answering and translation through multilingual answer targets, with 5.76 million cross-language QA records embedded in a 30.15-million-record intermediate mixture. On a 36,572-example synthetic test set, the final system improves all reported aggregate metrics over public Gemma 3 4B IT and 12B IT, including chrF++ gains in every direction. This study establishes a reproducible translation-focused adaptation recipe and a direction-level account of its strengths and remaining weaknesses.

Limitations

The test sources and references are synthetic, and the judge belongs to the same teacher-service family. Reference similarity and judge preferences may favor teacher-like outputs. The results therefore do not establish quality on human-authored benchmarks or out-of-domain text. The record-level split allows topic overlap; two source strings overlap after whitespace normalization, and the audit does not establish absence of semantic duplicates or pretraining exposure. Archived inference prompts differ across systems, so adaptation and prompting are not separated. Some baseline requests encountered context limits; their empty responses remain in the evaluation. COMET truncates long inputs, limiting the text considered by that metric. The judge subset contains only ten examples per direction. No human assessment, controlled mixture ablation, evaluation of intermediate checkpoints, or held-out QA evaluation is available. Training data have uneven language coverage, empty fields, and 1,961 conflicting answer-language directives. Internal validation sets are too small for reliable model selection. These limitations constrain both causal attribution and generalization of the reported gains.

Ethical Considerations

Wikipedia-derived and synthetic data can reflect uneven representation across languages and topics. Generated translations may alter names, factual claims, or register; the model has not been validated for high-stakes translation. Data and model use must respect applicable Wikipedia attribution/share-alike obligations, upstream dataset and teacher-service terms, and Gemma terms. The archived exports do not recover complete article-level provenance, so this paper does not assert a blanket license over all text. We report existing training and generation artifacts without collecting new human annotations. Full directional reporting and retention of empty outputs help make performance differences across languages visible.

References

Iman Barati, Arash Ghafouri, and Behrouz Minaei-Bidgoli. 2026. XHotpotQA: A benchmark for cross-lingual knowledge composition in multi-hop question answering. *Preprint*, arXiv:2608.27481.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. DeepSeek-V3.1: Model card. Hugging Face model repository. Accessed 12 September 2026.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hugging Face. 2025. PEFT: Parameter-efficient fine-tuning. Software, version 0.17.1.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. 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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Geyu Lin, Bin Wang, Zhengyuan Liu, and Nancy F. Chen. 2025. CrossIn: An efficient instruction tuning approach for cross-lingual knowledge alignment. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 12–23, Abu Dhabi. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Unbabel. 2022. wmt22-comet-da: Model card. Hugging Face model repository. Revision 2760a223ac95.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr

Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

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, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Continued-pretraining corpus

The five distinct source files registered for phase 1 contain the raw text rows in Table 4. Four Persian registrations refer to the same source file, so its rows are counted once in this table. The counts were measured from the available raw exports; they are neither unique-article counts nor the number of processed training sequences. The recorded training run reports 3,399,664 examples after preprocessing. No claim is made that the raw row sum alone reconstructs its tokenization, packing, or source repetition.

Language	Raw text rows
English	6,406,711
Persian	802,549
Arabic	1,216,238
Chinese	1,354,407
Egyptian Arabic	1,615,156
Total across distinct source files	11,395,061

Table 4: Available phase-1 Wikipedia sources before training preprocessing. The registered mixture repeats the Persian source four times.

B Construction and inference details

B.1 Multilingual answer inventory

The phase-2 inventory begins with three question formats (general, standard, and title-derived QA). Records are pooled by format and task in the released dataset. Their combined initial row count is 2,025,730. Gemma 3 27B IT (Gemma Team, 2025) generates Persian-to-English and Persian-to-Arabic variants, followed by English-to-Chinese, English-to-Hebrew, and English-to-Spanish variants. Thus part of the inventory is English-pivoted. Equivalence of all generated answer variants has not been exhaustively checked by human annotators. Wikipedia context supports upstream example construction; the final training record supplies only an instruction, input text, and target response.

Question columns supply QA inputs and answer columns supply QA targets. For answer-text MT, both columns contain answer variants. Sampling fractions for QA depend on the source group, direction, and response variant. For answer-text MT, the fraction of available content retained for a direction is

$$f(s,t) = \begin{cases} .70, & \{s,t\} = \{\text{fa},\text{en}\}, \\ .25, & \{s,t\} \cap \{\text{fa},\text{en}\} \neq \varnothing, \\ .08, & \text{otherwise.} \end{cases} \quad (6)$$

Here s, t are input and target languages and $s \neq t$. The first matching case applies. These fractions determine dataset construction; all retained answer-translation datasets are subsequently selected for phase 2. The construction records contain 10,695,112 MT rows; null filtering leaves 10,695,072. Across all tasks, summing retained construction rows reproduces 30,153,404 training records. The public release organizes these into 14 unified task configurations without dropping or resampling examples. Their internal split contains 30,153,102 training and 302 validation examples.

B.2 System-instruction composition

For cross-language complete-answer QA, the constructor samples a completeness directive and an answer-language directive, then combines them in randomized order with a sampled separator. Some same-language examples use only the completeness directive. Style-conditioned examples may add a formal or informal directive. The language in which a sentence is written is distinct from the answer language it requests. The construction uses random seed 42; this value controls reproducibility of sampling choices, not an additional modeling assumption. Prompt pools retain duplicate rows and their sampling mass.

B.3 Synthetic translation generation

Source generation uses a three-level topic taxonomy and attributes for genre, register, style, complexity, length, and optional linguistic/noise conditions. Stored metadata identifies DeepSeek-V3.1 (DeepSeek-AI, 2025); source generation uses temperature 0.5 and top-$p = 0.9$. The translation service is addressed through `deepseek-chat`, without a recoverable immutable serving revision for every call. A different target language is selected for each source. Translation generation uses temperature zero and a 5,000-token completion budget. The teacher is asked to preserve content, names, numbers, formatting, and register; these instructions do not guarantee translation quality.

The aggregate contains 365,718 pairs. A shuffled 90/10 record split produces 329,146 raw train and 36,572 test pairs. No split seed is recorded. Language-label normalization, pair deduplication, and removal of empty/equal texts yield 329,137 forward and 329,137 reverse training records. The net reduction is nine underlying pairs; a per-filter attribution is unavailable. The final selected total is 658,274, with an internal train/validation split of 658,267/7. The stored split memberships are preserved throughout the retrospective evaluation.

B.4 Archived inference

All three systems use temperature-zero decoding with a 5,000-token completion budget. HEXABEE-4B receives a short English direction instruction. The public Gemma models receive a longer English instruction specifying translation and formatting requirements, with quoted source text. The comparison therefore includes prompt differences. The full test contains 31 empty responses for Gemma 3 4B IT and none for the other two systems. Context-limit failures appear in the logs, so the empty responses cannot all be interpreted as linguistic failures. Empty strings and unsolicited wrappers remain in metric inputs.

B.5 Judge rubric and paired intervals

The reference-assisted judge sees source, candidate, and reference; the reference-free judge sees source and candidate. Both modes use a custom holistic 0–100 score and eight component scores, following the general model-based evaluation setting of Kocmi and Federmann (2023b,a). This is not a claim to reproduce GEMBA's exact protocol. Prompts request score caps of 10 for wrong-language output or untranslated copying, 40 for key mistranslations or hallucinations, and 60 for critical entity/number corruption. These are requested constraints, not verified error labels.

Let $J_{d,i}^{m,v}$ be the holistic judge score for system m, direction $d \in \Omega$, item $i \in \{1, \ldots, 10\}$, and view $v \in \{\text{ref}, \text{noref}\}$. For HEXABEE-4B (system h) and baseline b, the estimated score difference is

$$\widehat{\Delta}_v = \frac{1}{30} \sum_{d \in \Omega} \frac{1}{10} \sum_{i=1}^{10} (J_{d,i}^{h,v} - J_{d,i}^{b,v}). \qquad (7)$$

Each bootstrap replicate samples ten item indices with replacement within each direction, using the same sampled indices for both systems. We generate 10,000 replicates with seed 20260912 and use their 2.5th and 97.5th percentiles. These are paired, direction-stratified item-sampling intervals (Koehn, 2004).

C Phase-2 corpus composition

The following counts pool all retained records by question format and task. Summing construction records with non-null training fields exactly reproduces the training mixture. The released data have 14 unified task configurations; merging source partitions changes only packaging, not the examples. Source/target labels for answering tasks are independently checked against the notebook column policies. Counts describe records, not unique articles or independent question concepts.

Construction quality audit. The merger removes null fields but preserves empty strings. The final exports contain 6,742 empty input fields and 6,043 empty response fields (field counts

Subcorpus	Complete QA	Moderate QA	Formal QA	Informal QA	Translation	Total
General QA	3,952,557	208,364	330,194	284,650	2,481,368	7,257,133
Standard QA	7,240,185	300,875	1,540,222	411,033	5,554,592	15,046,907
Title QA	3,835,308	0	646,570	708,374	2,659,112	7,849,364

Table 5: All selected phase-2 data by question format and task, with source partitions pooled. A zero denotes an absent task export.

$s\backslash t$	fa	en	ar	zh	he	es
fa	2,025,720	764,747	202,573	101,287	19,744	1,139
en	1,418,003	2,025,730	101,287	462,076	39,489	0
ar	405,143	202,573	2,025,730	20,258	4,936	0
zh	202,572	101,287	40,514	2,025,730	4,114	822
he	49,362	29,616	9,873	4,936	987,223	0
es	107,249	76,607	15,322	15,322	4,936	1,532,130

Table 6: Complete QA records by input language (rows) and response language (columns). Diagonal cells are same-language answering; translation excludes identity directions.

$s\backslash t$	fa	en	ar	zh	he	es
fa	4,545,437	931,711	202,573	101,287	19,744	1,139
en	2,513,228	2,582,276	101,287	462,076	39,489	0
ar	460,798	202,573	2,025,730	20,258	4,936	0
zh	222,051	101,287	40,514	2,025,730	4,114	822
he	57,710	29,616	9,873	4,936	987,223	0
es	115,597	76,607	15,322	15,322	4,936	1,532,130

Table 7: All QA response variants (complete, moderate, formal, informal) records by input language (rows) and response language (columns). Diagonal cells are same-language answering; translation excludes identity directions.

can overlap within a record); no system field is empty. These entries are included in the recorded export counts. The raw construction-to-export reduction is 69 records, including 40 translation records. Two complete-answer exports contain 822 Chinese-to-Spanish and 1,139 Persian-to-Spanish records whose language-directive pools request English. Both conflicts are verified in stored examples. They account for all 1,961 cross-language Spanish-target complete-answer records. Consequently Spanish cross-language answering with a correctly specified target is not established by this subset, although Spanish is present in same-language QA and translation supervision. The historical model is evaluated as trained; the audit does not retroactively repair its supervision.

D Prompt sampling details

For instruction component g (such as answer language or extent), let $\mathcal{P}_{g,p,t}$ be the list of sentences written in language p for target language t. Let $\mathcal{A}_{g,t}$ contain the languages with available,

$s\backslash t$	fa	en	ar	zh	he	es
fa	0	1,418,003	506,429	506,429	246,806	383,029
en	1,418,003	0	506,433	506,433	246,806	383,033
ar	506,429	506,433	0	162,059	78,978	122,571
zh	506,429	506,433	162,059	0	78,978	122,571
he	246,806	246,806	78,978	78,978	0	78,978
es	383,029	383,033	122,571	122,571	78,978	0

Table 8: Answer-text MT records by input language (rows) and response language (columns). Diagonal cells are same-language answering; translation excludes identity directions.

	Answer-language		Completeness	
Prompt language	Rows	Unique	Rows	Unique
fa	1302	1110	1076	836
en	1302	944	1000	759
ar	1060	817	1032	792
zh	976	755	1024	784
he	914	691	1080	840
es	–	–	1009	728

Table 9: Representative prompt pools before record sampling. Counts use the pool whose target-language label matches its prompt language; completeness text does not itself specify the answer language. Unique counts deduplicate exact sentence strings. Repeated rows remain in the sampler. A dash indicates no such pool.

nonempty pools. The builder normalizes configured nonnegative weights $w_{g,p}$ over $\mathcal{A}_{g,t}$, chooses p, and samples row index j uniformly:

$$P(p, j \mid g, t) = \frac{w_{g,p}}{\sum_{u \in \mathcal{A}_{g,t}} w_{g,u}} \frac{1}{|\mathcal{P}_{g,p,t}|}. \quad (8)$$

Here $p \in \mathcal{A}_{g,t}$ and $1 \le j \le |\mathcal{P}_{g,p,t}|$. The configured available weights have a positive sum. Repeated strings in a list therefore receive repeated sampling mass; uniformity is over rows, not distinct paraphrases. The prompt-generation scripts expand seed sentences with model-generated paraphrases, combine outputs, and substitute localized language names into target placeholders. Appendix C reports pool sizes and exact duplicate counts. Changing p changes how a directive is expressed; changing t changes the required response. Our core supervision concerns the latter.

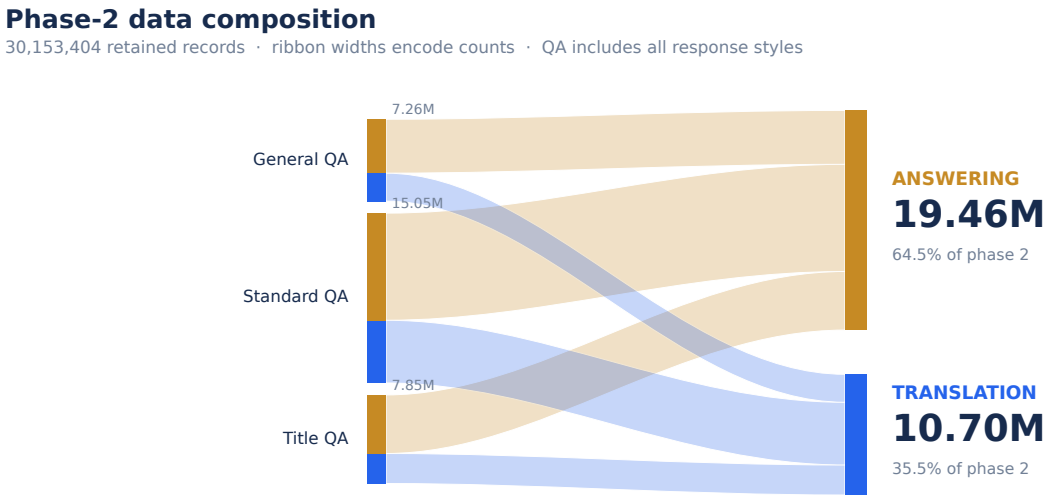


Figure 5: Retained phase-2 records flowing from the three pooled question formats to answering and translation tasks. Ribbon widths encode record counts. Translation and several answering variants are jointly selected; source-row reuse across these tasks means the plotted counts are not independent documents.

D.1 Final-phase translation instructions

For each direction, the constructor chooses a prompt-language pool, then uniformly samples a sentence from that pool. Its principal weights are

$$\pi(p \mid s,t) = \begin{cases} .55, & p = t, \\ .30, & p = s, \\ w_p(s,t), & p \notin \{s,t\}, \end{cases} \tag{9}$$

The residual satisfies $\sum_{p \notin \{s,t\}} w_p = .15$. Remaining weights use a script-based heuristic: Persian/Arabic/Hebrew form an RTL group, English/Spanish a Latin group, and Chinese a CJK group. These are implementation groups, not genealogical language families. The allocation below specifies these weights. Empty pools are excluded and available weights renormalized by the sampler. Seed 42 controls the sampled construction choices; it does not retroactively seed source generation or the earlier raw split.

A final-phase translation record has the training schema `system`, `instruct`, `output`. Per-direction intermediate files retain explicit source/target labels, but the merger keeps the three training fields, drops missing values, and escapes media-like tags as text. Here p denotes the language of the sampled translation instruction and supplies surface-form diversity. It does not add a loss that explicitly equates outputs across prompt languages, and our evaluation does not estimate such invariance.

Let $U = \mathcal{L} \setminus \{s,t\}$, $M \subseteq U$ contain languages sharing an implementation script group with s or t, and $O = U \setminus M$. The constructor reserves 0.55 for t and 0.30 for s. With residual $R = .15$, define

$$g = \min(R/2, .01|O|), \tag{10}$$

$$h = \begin{cases} \min(.10, (R-g)/|M|), & |M| > 0, \\ 0, & |M| = 0, \end{cases} \tag{11}$$

$$r_0 = R - g - |M|h. \tag{12}$$

For $p \in U$, the residual weight is

$$w_p = \mathbf{1}[p \in O]\frac{g}{|O|} + \mathbf{1}[p \in M]h + \frac{r_0}{|U|}, \tag{13}$$

where a fraction on an empty group contributes zero. This matches the implementation's residual redistribution before rounding and final sampler normalization. For Persian-to-English, the unrounded weights in order (fa,en,es,he,zh,ar) are $(.30, .55, .046\overline{6}, .046\overline{6}, .01, .046\overline{6})$. Here $\mathbf{1}[\cdot]$ is the indicator function. With $\mathcal{P}_{p,s,t}$ the nonempty prompt pool, the probability of selecting row j from language pool p is

$$P(j, p \mid s,t) = \frac{\pi(p \mid s,t)}{|\mathcal{P}_{p,s,t}|}, \quad 1 \le j \le |\mathcal{P}_{p,s,t}|. \tag{14}$$

If pools are unavailable, π is renormalized over the available pools. The six-language assignment

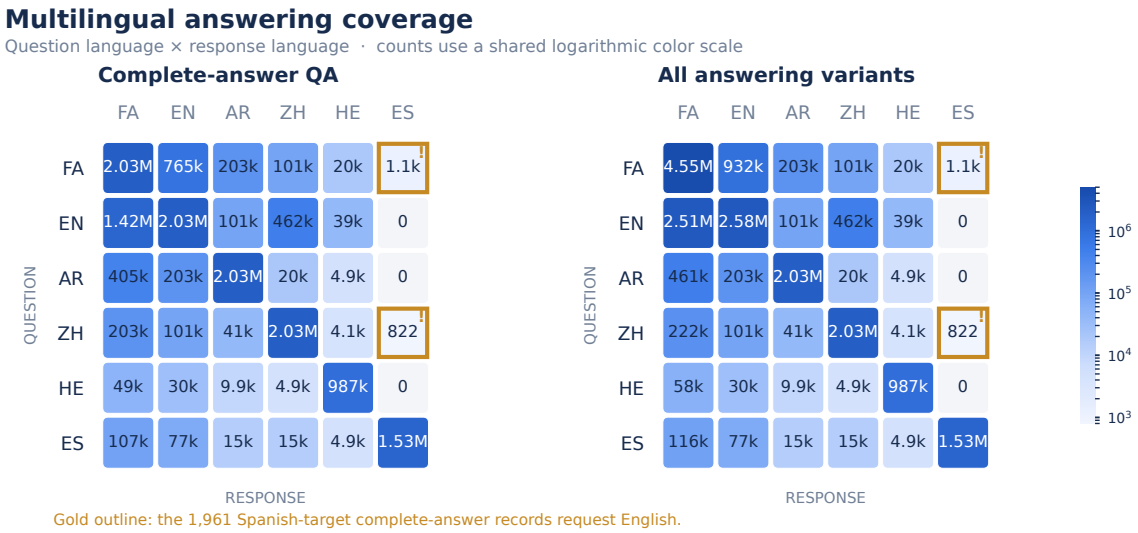


Figure 6: Realized question/answer language coverage. Left: complete-answer QA. Right: all answering variants. Color uses a logarithmic count scale. Gold outlines mark the two Spanish-target cells whose complete-answer directives incorrectly request English (1,961 records). Zero cells indicate absent exported combinations.

has full theoretical support, but observed coverage must be measured from instantiated prompts rather than inferred from the formula.

D.2 Partition audit

Streaming comparison of raw source strings finds zero exact train–test source overlaps. After collapsing whitespace, two test source texts match training sources, corresponding to 0.0055% of the test partition; neither occurs in the 300-item judge subset. This is a narrow source-text audit: it does not prove absence of semantic duplicates, topic overlap, overlap with reverse inputs, or exposure during pretraining. The split is by record rather than by topic or Wikipedia article. We retain this distinction when interpreting generalization.

E Training and serialization

The final stage uses PEFT 0.17.1 (Hugging Face, 2025), Transformers 4.55.4 (Wolf et al., 2020), PyTorch 2.7.1+cu126, Datasets 3.6.0, Tokenizers 0.21.1, and DeepSpeed 0.16.9 with ZeRO stage 0. AdamW (Loshchilov and Hutter, 2019) parameters are $\beta_1 = .9$, $\beta_2 = .999$, and $\epsilon = 10^{-8}$. Training uses 4,096-token sequences, response-only supervision, checkpoints every 600 steps, and internal evaluation every 200 steps. The final adapter has no bias adaptation, DoRA, or rank-stabilized LoRA. The saved configuration of the continued cross-lingual adapter also uses dropout .005 despite a different value in the nearby YAML.

For a text-only example, the serialized prefix is shown schematically below; literal line breaks separate fields:

```
<bos><start_of_turn>user
[system instruction]

[source text]<end_of_turn>
<start_of_turn>model
[target text]<end_of_turn>
```

$$W'_j = W_j + \frac{\alpha}{r} B_j A_j, \tag{15}$$

where $W_j \in \mathbb{R}^{d_{o,j} \times d_{i,j}}$ is projection j with input dimension $d_{i,j}$ and output dimension $d_{o,j}$, $A_j \in \mathbb{R}^{r \times d_{i,j}}$ and $B_j \in \mathbb{R}^{d_{o,j} \times r}$ are trainable factors, r is adapter rank, and α is its scale (Hu et al., 2022). The saved adapters use $r = 256$, $\alpha = 512$, and dropout 0.005. These standard LoRA settings are implementation choices; the method's supervision relations are given in the main text.

The model's 34 text layers have hidden size 2,560, eight query heads, four key/value heads, and MLP intermediate size 10,240. The multimodal checkpoint retains its vision components, although final-stage adaptation affects only text projections. The nominal 4B size identifies the Gemma model family. For the adapted projections, the parameter count is

$$N_{\text{LoRA}} = r \sum_j (d_{i,j} + d_{o,j}) = 476{,}839{,}936. \tag{16}$$

Each layer contributes two $2{,}560 \times 2{,}048$ attention projections, two $2{,}560 \times 1{,}024$ projections, and

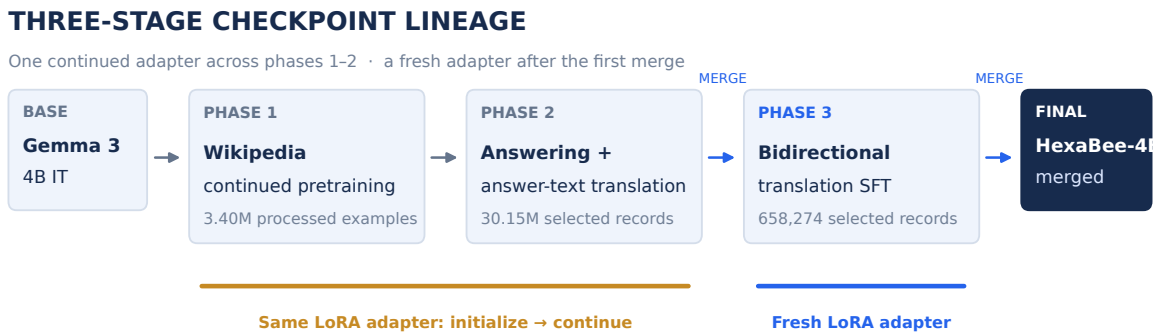


Figure 7: Training lineage. The translation stage starts a new adapter on the merged cross-lingual model. Phase 2 combines system-directed answering with translations of multilingual answer texts; phase 3 specializes the merged model on independent synthetic translations.

Setting	Wikipedia adaptation	Cross-lingual SFT	Translation SFT
Training examples	3,399,664	30,153,102	658,267
Internal validation examples	Not established	302	7
Epochs / optimizer steps	1 / 13,280	1 / 6,544	1 / 3,429
CUDA processes	4	6	2
Per-device batch / accumulation	16 / 4	12 / 64	12 / 8
Effective batch	256	4,608	192
Peak learning rate	10^{-4}	10^{-4}	10^{-4}
Schedule / warmup fraction	cosine / .05	cosine / .05	cosine / .05
Seed	111	42	42
Runtime (hours)	59.99	306.90	27.45
Adapter handling	new on public base	continue previous	new after merge

Table 10: Executed settings from training logs and generated model cards. Counts have stage-specific semantics and should not be added as unique documents. Current nearby YAMLs may describe other runs; logged values take precedence.

three $2{,}560 \times 10{,}240$ MLP projections, up to transposition. Summing their input/output dimensions over 34 layers verifies the adapter count.

F Metrics and results by language

Language codes are ar (Arabic), zh (Chinese), en (English), he (Hebrew), fa (Persian), and es (Spanish). Direction tables keep source and target distinct. Target summaries give each of the five incoming directions equal weight; they are direction-macro summaries, not recomputed pooled corpora. Judge dimension scores remain on a 0–100 scale and are not calibrated accuracy percentages. All raw model responses are retained for scoring, including quotation wrappers. The machine-readable artifacts include metric signatures and sample counts. Writing $M_{s,t}$ for the score of direction (s, t), the target-language summary is

$$M_{\to t} = \frac{1}{5} \sum_{s \in \mathcal{L} \setminus \{t\}} M_{s,t}. \tag{17}$$

Pooled corpus chrF++ is 60.77 for Gemma 3 4B IT, 71.13 for Gemma 3 12B IT, and 77.30 for HexaBee-4B. These differ from the direction-macro values in Table 2. Let $\mathcal{G}$ index the configured n-gram types, each defined by a unit (character or word) and an order. For $g \in \mathcal{G}$, let H_g, R_g, and C_g denote corpus hypothesis, reference, and matched counts. With $E = \{g \in \mathcal{G} : H_g > 0, R_g > 0\}$, the effective-order averages are

$$P = \frac{1}{|E|} \sum_{g \in E} \frac{C_g}{H_g}, \qquad R = \frac{1}{|E|} \sum_{g \in E} \frac{C_g}{R_g}. \tag{18}$$

For chrF++, orders include character 1–6 and word 1–2; chrF uses only the character orders. SacreBLEU then computes

$$\mathrm{chrF}_\beta = 100(1 + \beta^2) \frac{PR}{\beta^2 P + R}, \quad \beta = 2, \tag{19}$$

with a zero score for an empty effective set or zero denominator. It aggregates corpus counts before the harmonic combination, rather than averaging sentence scores. BLEU is computed from

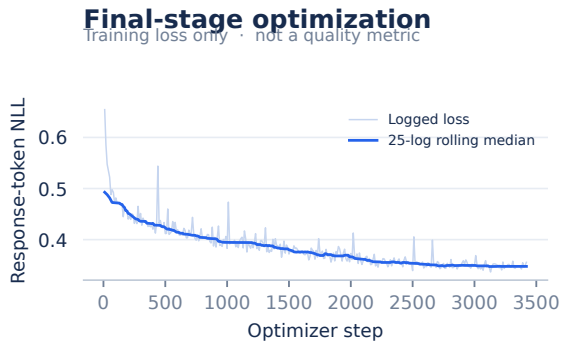


Figure 8: Recorded response-token training loss during final translation adaptation, with a centered rolling median over up to 25 logged values for visual guidance. This is an optimization diagnostic, not a translation-quality or generalization metric.

clipped corpus precisions p_n and total hypothesis/reference lengths c, r:

$$\mathrm{BLEU} = 100\,\mathrm{BP}\exp\left(\frac{1}{4}\sum_{n=1}^{4}\log p_n\right), \quad (20)$$

where $\mathrm{BP} = \min(1, \exp(1 - r/c))$ for $c > 0$, with SacreBLEU's exponential smoothing for zero matches. TER uses the Tercom-style edit search:

$$\mathrm{TER} = 100\frac{N_{\mathrm{sub}} + N_{\mathrm{del}} + N_{\mathrm{ins}} + N_{\mathrm{shift}}}{N_{\mathrm{ref}}}. \quad (21)$$

Here N_{sub}, N_{del}, N_{ins}, and N_{shift} count substitutions, deletions, insertions, and shifts; N_{ref} is the total reference-token count. This search approximates the edit sequence with shifts; the score should not be described as a certified global minimum of all possible edit programs. Corpus and macro aggregation are kept distinct for both metrics.

Let f_ω be the COMET regression model with fixed checkpoint parameters ω and τ_{512} its tokenizer/encoder truncation operation. For N examples, let x_i, $\hat{y}_i$, and y_i denote source, candidate, and reference text, respectively. The scaled corpus mean is

$$\frac{100}{N}\sum_{i=1}^{N} f_\omega(\tau_{512}(x_i), \tau_{512}(\hat{y}_i), \tau_{512}(y_i)). \quad (22)$$

Direction-macro COMET instead averages the 30 direction means. Scaling by 100 does not turn the regression output into an accuracy percentage.

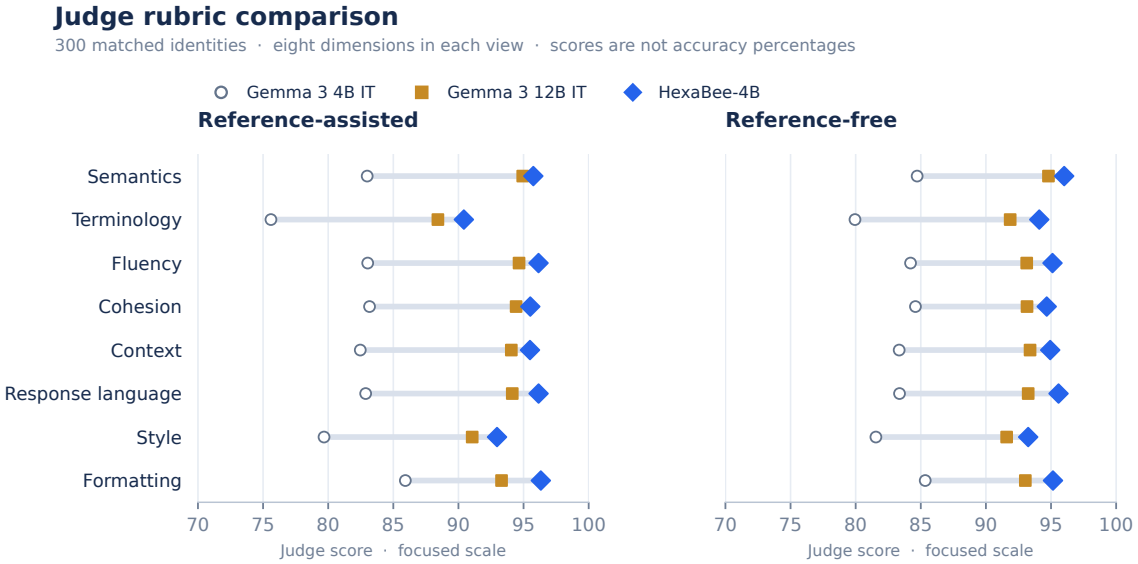


Figure 9: All eight judge dimensions, averaged over the same 300 identities. Both views compare the same three systems; they remain judgments from one evaluator and are not independent human assessments.

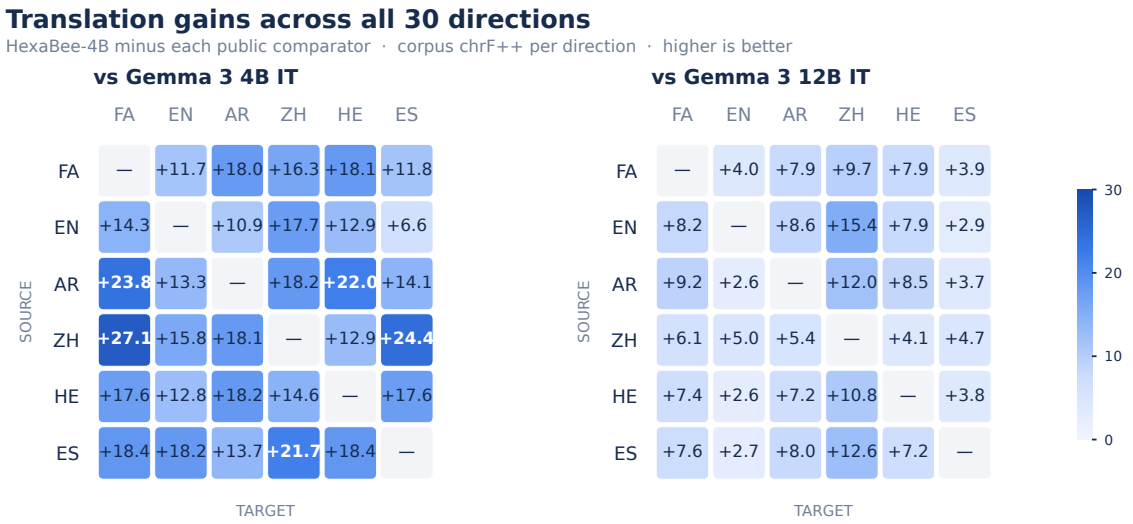


Figure 10: Full-test chrF++ gains of HEXABEE-4B over both public comparison systems across all 30 source–target directions.

Target	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.
All	4B IT	36,572	40.70	58.91	57.11	51.40	85.88	80.32	81.43
	12B IT	36,572	50.54	68.70	66.83	38.89	88.20	91.73	91.97
	HEXABEE-4B	36,572	**62.03**	**75.08**	**73.75**	**30.05**	**89.86**	**92.92**	**94.10**
ar	4B IT	6,120	34.09	62.25	59.33	56.99	86.64	82.56	81.60
	12B IT	6,120	43.14	70.43	67.69	44.60	88.52	89.12	89.78
	HEXABEE-4B	6,120	**56.17**	**77.09**	**75.10**	**33.62**	**90.17**	**92.72**	**93.60**
zh	4B IT	6,135	35.22	31.87	30.97	60.19	84.52	75.22	76.00
	12B IT	6,135	40.15	38.40	36.57	54.28	87.01	89.92	90.82
	HEXABEE-4B	6,135	**54.78**	**50.31**	**48.67**	**38.68**	**90.00**	**91.82**	**92.60**
en	4B IT	6,010	55.28	74.13	72.40	36.35	87.40	86.26	86.84
	12B IT	6,010	67.16	84.77	83.36	22.59	88.90	**94.78**	94.42
	HEXABEE-4B	6,010	**75.06**	**87.68**	**86.76**	**17.00**	**89.83**	93.52	**94.66**
he	4B IT	6,196	31.94	54.98	53.02	59.27	85.87	76.00	78.44
	12B IT	6,196	43.32	64.53	62.78	46.14	88.17	91.08	91.40
	HEXABEE-4B	6,196	**54.71**	**70.90**	**69.90**	**42.17**	**89.68**	**93.16**	**94.30**
fa	4B IT	6,069	32.44	57.57	55.67	59.16	84.81	76.86	77.86
	12B IT	6,069	43.61	70.24	68.22	41.59	88.32	91.04	92.60
	HEXABEE-4B	6,069	**57.43**	**77.34**	**75.90**	**31.68**	**89.91**	**92.06**	**94.34**
es	4B IT	6,042	55.21	72.69	71.24	36.44	86.02	85.04	87.86
	12B IT	6,042	65.85	83.83	82.35	24.13	88.29	**94.44**	92.82
	HEXABEE-4B	6,042	**74.04**	**87.18**	**86.16**	**17.13**	**89.57**	94.24	**95.12**

Table 11: Overall and target-language comparisons. Automatic metrics are direction-macro averages on the full test set; holistic judge means use 300 matched items overall or 50 per target. COMET is scaled by 100; TER is lower-is-better. Bold marks the best displayed score and underlining the second distinct displayed score within each group; ties share rank. The adapted model is tinted for easy comparison. These recorded systems use different inference prompts.

Direction	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.	chrF++_{10}	TER_{10} ↓	COMET_{10}
$zh \to ar$	4B IT	1229	23.68	51.49	48.28	80.63	84.07	72.60	71.60	37.49	100.59	80.35
	12B IT	1229	32.53	64.44	60.91	61.20	87.91	**91.60**	92.50	62.38	53.76	87.21
	HEXABEE-4B	1229	**42.46**	**69.06**	**66.34**	**57.30**	**89.15**	90.40	**93.00**	**71.36**	**38.78**	**88.68**
$en \to ar$	4B IT	1213	37.50	66.77	63.77	51.70	87.99	86.50	85.20	65.44	44.06	88.04
	12B IT	1213	42.30	68.56	66.09	47.70	87.47	81.10	82.20	57.25	48.09	85.10
	HEXABEE-4B	1213	**55.60**	**76.61**	**74.66**	**35.72**	**90.06**	**93.40**	**94.10**	**76.86**	**27.55**	**90.48**
$he \to ar$	4B IT	1230	39.08	63.41	61.23	48.02	86.42	78.30	74.70	50.31	52.36	80.29
	12B IT	1230	51.01	74.60	72.25	32.19	89.30	88.40	87.90	64.18	38.88	84.90
	HEXABEE-4B	1230	**63.65**	**80.99**	**79.44**	**23.26**	**90.59**	**92.60**	**93.60**	**69.00**	**31.11**	**86.10**
$fa \to ar$	4B IT	1228	32.88	63.15	60.05	56.62	87.51	88.70	89.80	68.97	38.75	89.71
	12B IT	1228	46.62	72.69	70.14	38.53	89.75	92.60	93.40	73.76	31.76	90.41
	HEXABEE-4B	1228	**61.00**	**79.59**	**78.00**	**24.43**	**91.27**	**94.40**	**95.10**	**81.87**	**20.81**	**91.55**
$es \to ar$	4B IT	1220	37.28	66.41	63.33	47.96	87.21	86.70	86.70	64.12	44.99	88.36
	12B IT	1220	43.22	71.87	69.06	43.37	88.15	91.90	**92.90**	69.86	37.83	88.93
	HEXABEE-4B	1220	**58.14**	**79.19**	**77.07**	**27.40**	**89.78**	**92.80**	92.20	**76.58**	**27.15**	**89.85**

Table 12: All systems translating into Arabic. Columns through COMET use the full direction-specific test set (N). The last 5 columns use the ten shared judge-subset items per direction. Bold/underline mark first/second within the same direction and metric; TER is lower-is-better.

Direction	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.	chrF++$_{10}$	TER$_{10}$ ↓	COMET$_{10}$
$ar \to zh$	4B IT	1201	36.56	32.17	30.93	58.04	83.81	73.60	73.90	22.92	62.38	81.89
	12B IT	1201	42.36	39.84	37.13	50.06	86.94	91.80	91.90	30.13	51.75	87.06
	HEXABEE-4B	1201	**57.09**	**51.82**	**49.12**	**35.01**	**89.69**	**93.70**	**94.20**	**40.72**	**35.73**	**89.56**
$en \to zh$	4B IT	1276	30.42	27.90	27.56	69.42	83.44	82.30	82.80	29.76	49.49	83.68
	12B IT	1276	32.77	31.48	29.94	67.91	85.07	91.80	93.50	30.05	44.73	87.02
	HEXABEE-4B	1276	**50.33**	**45.82**	**45.30**	**45.39**	**90.02**	**93.00**	**93.60**	**44.30**	**31.80**	**89.95**
$he \to zh$	4B IT	1202	41.02	38.25	37.22	51.04	86.17	76.10	77.20	29.02	66.88	84.29
	12B IT	1202	44.57	42.44	40.97	45.95	87.67	93.00	93.50	38.94	52.52	87.92
	HEXABEE-4B	1202	**58.14**	**53.41**	**51.80**	**33.49**	**89.75**	**94.20**	**94.30**	**50.28**	**34.08**	**90.12**
$fa \to zh$	4B IT	1251	38.95	35.74	34.01	53.12	86.19	74.70	75.00	37.67	48.16	83.33
	12B IT	1251	45.23	42.86	40.58	44.42	88.67	92.10	90.50	46.31	34.15	89.23
	HEXABEE-4B	1251	**57.23**	**52.86**	**50.30**	**34.91**	**90.41**	**94.40**	**94.60**	**58.36**	**23.84**	**91.03**
$es \to zh$	4B IT	1205	29.14	25.27	25.15	69.33	83.01	69.40	71.10	13.03	71.67	79.39
	12B IT	1205	35.84	35.37	34.24	63.07	86.73	80.90	84.70	17.67	62.46	82.49
	HEXABEE-4B	1205	**51.11**	**47.63**	**46.84**	**44.61**	**90.14**	**83.80**	**86.30**	**24.63**	**44.99**	**85.80**

Table 13: All systems translating into Chinese. Columns through COMET use the full direction-specific test set (N). The last 5 columns use the ten shared judge-subset items per direction. Bold/underline mark first/second within the same direction and metric; TER is lower-is-better.

Direction	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.	chrF++$_{10}$	TER$_{10}$ ↓	COMET$_{10}$
$ar \to en$	4B IT	1200	59.91	75.60	74.32	29.38	86.98	82.00	83.30	64.16	41.19	85.14
	12B IT	1200	70.24	86.13	84.97	21.03	88.74	**92.40**	**94.80**	81.37	23.10	88.22
	HEXABEE-4B	1200	**76.36**	**88.32**	**87.58**	**16.02**	**89.58**	90.10	94.60	**85.46**	**16.95**	**89.33**
$zh \to en$	4B IT	1213	38.89	64.46	62.03	59.53	86.20	64.40	65.90	53.20	73.92	81.00
	12B IT	1213	47.49	75.48	72.87	40.65	87.93	**95.70**	94.90	73.06	39.17	88.17
	HEXABEE-4B	1213	**58.71**	**79.80**	**77.87**	**32.67**	**89.16**	95.20	**95.60**	**78.02**	**30.94**	**89.30**
$he \to en$	4B IT	1185	63.49	77.80	76.72	28.99	87.63	**93.40**	**92.40**	81.10	21.63	87.26
	12B IT	1185	73.72	87.81	86.82	16.72	89.22	91.80	89.20	86.07	14.94	87.78
	HEXABEE-4B	1185	**79.87**	**90.07**	**89.47**	**12.54**	**90.03**	90.80	90.00	**89.03**	**10.89**	**88.69**
$fa \to en$	4B IT	1234	57.78	76.96	75.53	29.13	88.46	95.10	95.40	77.19	27.66	87.40
	12B IT	1234	67.46	84.48	83.21	20.09	89.29	96.60	95.80	80.82	22.40	87.88
	HEXABEE-4B	1234	**76.31**	**88.01**	**87.22**	**14.41**	**90.16**	**96.70**	**96.70**	**85.43**	**16.07**	**88.65**
$es \to en$	4B IT	1178	56.34	75.84	73.42	34.73	87.71	96.40	97.20	86.98	16.42	89.07
	12B IT	1178	76.88	89.96	88.91	14.43	89.31	**97.40**	**97.40**	88.29	13.85	89.11
	HEXABEE-4B	1178	**84.06**	**92.21**	**91.63**	**9.37**	**90.24**	94.80	96.40	**91.34**	**9.19**	**89.94**

Table 14: All systems translating into English. Columns through COMET use the full direction-specific test set (N). The last 5 columns use the ten shared judge-subset items per direction. Bold/underline mark first/second within the same direction and metric; TER is lower-is-better.

Direction	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.	chrF++$_{10}$	TER$_{10}$ ↓	COMET$_{10}$
$ar \to he$	4B IT	1255	33.86	55.36	53.52	55.20	85.90	70.40	66.30	48.01	52.39	84.15
	12B IT	1255	48.39	68.91	67.01	37.34	88.91	91.70	93.20	70.71	30.88	88.60
	HEXABEE-4B	1255	**61.41**	**76.53**	**75.50**	**28.87**	**91.03**	**93.00**	**94.70**	**78.93**	**20.92**	**90.49**
$zh \to he$	4B IT	1232	19.81	39.55	38.42	88.40	82.50	85.00	87.00	53.38	61.20	85.05
	12B IT	1232	27.88	48.10	47.27	**78.13**	85.34	90.30	88.60	58.72	53.58	86.56
	HEXABEE-4B	1232	**35.82**	**51.34**	**51.36**	90.09	**86.74**	**93.70**	**94.70**	**63.72**	**46.14**	**87.24**
$en \to he$	4B IT	1216	36.21	61.46	59.13	49.77	87.17	71.20	77.20	59.56	44.59	86.58
	12B IT	1216	44.79	65.92	64.17	41.62	88.31	92.60	92.10	64.07	38.82	88.78
	HEXABEE-4B	1216	**56.73**	**73.34**	**72.04**	**32.13**	**89.85**	**93.40**	**92.80**	**71.85**	**28.26**	**89.52**
$fa \to he$	4B IT	1238	33.83	58.93	56.54	52.35	87.53	70.40	75.60	54.20	51.63	87.48
	12B IT	1238	45.84	69.08	66.79	38.09	89.74	90.10	91.20	72.00	32.67	90.98
	HEXABEE-4B	1238	**58.38**	**76.09**	**74.67**	**33.36**	**90.94**	**92.10**	**94.50**	**79.33**	**23.28**	**92.35**
$es \to he$	4B IT	1255	35.99	59.59	57.50	50.63	86.26	83.00	86.10	65.78	53.33	88.65
	12B IT	1255	49.71	70.65	68.68	35.51	88.55	90.70	91.90	73.23	29.21	89.87
	HEXABEE-4B	1255	**61.20**	**77.18**	**75.92**	**26.40**	**89.85**	**93.60**	**94.80**	**80.26**	**21.41**	**90.87**

Table 15: All systems translating into Hebrew. Columns through COMET use the full direction-specific test set (N). The last 5 columns use the ten shared judge-subset items per direction. Bold/underline mark first/second within the same direction and metric; TER is lower-is-better.

Direction	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.	chrF++$_{10}$	TER$_{10}$ ↓	COMET$_{10}$
$ar \to fa$	4B IT	1228	32.96	57.46	55.29	56.93	84.60	73.20	75.10	58.50	48.14	84.77
	12B IT	1228	45.58	71.82	69.85	40.84	88.30	**89.80**	92.60	70.16	34.38	89.77
	HexaBee-4B	1228	**62.29**	**80.41**	**79.05**	**23.95**	**90.46**	89.50	**95.10**	**80.22**	**21.59**	**91.29**
$zh \to fa$	4B IT	1208	19.17	41.95	40.29	90.09	81.51	51.60	52.70	32.82	98.86	79.60
	12B IT	1208	32.94	63.68	61.27	55.33	87.85	89.40	91.50	63.25	49.32	89.01
	HexaBee-4B	1208	**45.05**	**69.11**	**67.38**	**55.17**	**89.04**	**91.00**	**93.10**	**69.82**	**36.28**	**90.41**
$en \to fa$	4B IT	1239	34.35	62.41	60.48	52.63	86.10	89.70	88.80	64.17	48.39	87.44
	12B IT	1239	42.64	68.48	66.60	41.52	88.24	91.70	91.90	67.09	42.60	88.90
	HexaBee-4B	1239	**56.25**	**76.24**	**74.81**	**29.15**	**90.13**	**93.00**	**93.80**	**76.46**	**27.69**	**90.36**
$he \to fa$	4B IT	1198	40.23	64.44	62.67	44.97	86.19	90.40	91.80	63.53	43.59	86.47
	12B IT	1198	50.15	74.71	72.89	33.55	88.98	93.40	94.80	72.82	32.88	88.59
	HexaBee-4B	1198	**63.73**	**81.50**	**80.26**	**23.03**	**90.18**	**94.50**	**95.90**	**78.81**	**22.59**	**89.28**
$es \to fa$	4B IT	1196	35.51	61.60	59.63	51.17	85.67	79.40	80.90	55.27	55.07	84.60
	12B IT	1196	46.74	72.49	70.47	36.71	88.22	90.90	92.20	65.24	43.73	88.41
	HexaBee-4B	1196	**59.82**	**79.44**	**78.03**	**27.11**	**89.71**	**92.30**	**93.80**	**74.01**	**31.84**	**90.33**

Table 16: All systems translating into Persian. Columns through COMET use the full direction-specific test set (N). The last 5 columns use the ten shared judge-subset items per direction. Bold/underline mark first/second within the same direction and metric; TER is lower-is-better.

Direction	System	N	BLEU	chrF	chrF++	TER↓	COMET	Ref. J.	No-ref. J.	chrF++$_{10}$	TER$_{10}$ ↓	COMET$_{10}$
$ar \to es$	4B IT	1257	59.24	74.94	73.59	30.24	85.60	88.20	91.10	82.38	20.28	87.14
	12B IT	1257	68.19	85.36	84.08	20.85	87.93	**94.10**	**96.30**	86.38	14.91	88.39
	HexaBee-4B	1257	**77.23**	**88.55**	**87.73**	**14.41**	**89.37**	91.80	94.20	**88.63**	**12.20**	**88.66**
$zh \to es$	4B IT	1173	32.40	55.60	53.72	68.49	83.08	81.90	86.20	63.96	53.96	86.22
	12B IT	1173	48.63	75.81	73.37	44.50	87.27	93.60	92.90	75.41	35.64	88.58
	HexaBee-4B	1173	**58.16**	**80.03**	**78.11**	**31.49**	**89.03**	**94.10**	**94.90**	**81.32**	**24.58**	**89.61**
$en \to es$	4B IT	1188	66.60	84.25	82.63	23.34	88.08	93.20	94.90	81.53	22.75	87.87
	12B IT	1188	73.09	87.64	86.37	17.24	88.72	**94.30**	95.90	84.14	19.54	87.99
	HexaBee-4B	1188	**79.33**	**90.12**	**89.27**	**12.45**	**89.61**	93.90	**96.20**	**86.92**	**14.98**	**88.84**
$he \to es$	4B IT	1183	58.27	71.77	70.79	30.88	85.57	78.40	77.20	58.37	49.60	82.44
	12B IT	1183	71.49	85.66	84.60	17.18	88.34	96.00	86.70	85.00	30.21	87.60
	HexaBee-4B	1183	**79.39**	**89.05**	**88.43**	**11.91**	**89.55**	**96.50**	**97.00**	**89.36**	**11.67**	**89.92**
$fa \to es$	4B IT	1241	59.54	76.89	75.45	29.25	87.78	83.50	89.90	77.75	27.47	86.37
	12B IT	1241	67.83	84.70	83.34	20.90	89.22	94.20	92.30	79.56	24.97	86.03
	HexaBee-4B	1241	**76.09**	**88.16**	**87.27**	**15.38**	**90.28**	**94.90**	**93.30**	**84.07**	**20.04**	**87.36**

Table 17: All systems translating into Spanish. Columns through COMET use the full direction-specific test set (N). The last 5 columns use the ten shared judge-subset items per direction. Bold/underline mark first/second within the same direction and metric; TER is lower-is-better.

		Reference-assisted judge								Reference-free judge							
Dir.	System	Sem.	Term.	Flu.	Coh.	Ctx.	Lang.	Style	Fmt.	Sem.	Term.	Flu.	Coh.	Ctx.	Lang.	Style	Fmt.
$zh \to ar$	4B IT	72.0	66.5	73.8	72.1	71.8	73.9	68.8	73.5	74.1	69.3	74.0	72.7	72.7	74.8	71.5	75.8
	12B IT	**94.8**	87.5	94.2	92.8	93.8	94.6	90.5	93.0	**95.1**	92.0	91.8	92.6	92.8	93.6	91.6	94.0
	HexaBee-4B	92.4	**89.3**	**95.2**	**93.3**	**93.9**	**94.8**	**91.3**	**96.0**	95.0	**92.1**	**94.3**	**93.8**	**94.5**	**95.0**	**92.9**	**95.0**
$en \to ar$	4B IT	91.0	80.2	90.2	91.0	90.1	89.2	85.5	91.5	88.0	79.7	88.5	88.6	86.6	89.0	85.0	91.5
	12B IT	83.3	76.0	94.3	94.5	82.8	84.5	81.3	95.0	94.0	88.5	93.4	92.2	93.2	84.0	91.2	94.5
	HexaBee-4B	**95.7**	**88.4**	**97.6**	**96.1**	**95.9**	**97.1**	**93.2**	**97.3**	**96.2**	**95.0**	**94.3**	**94.9**	**95.3**	**94.9**	**92.2**	**95.5**
$he \to ar$	4B IT	79.1	73.8	80.3	80.3	79.5	80.1	77.0	88.0	77.0	71.2	77.4	78.6	76.1	77.0	75.5	79.8
	12B IT	93.8	83.2	92.4	91.2	91.6	90.3	88.2	88.5	90.7	85.5	88.9	89.5	89.7	88.1	88.7	88.5
	HexaBee-4B	**96.3**	**89.7**	**96.0**	**94.7**	**92.8**	**95.6**	**92.0**	**96.0**	**95.3**	**93.7**	**94.8**	**94.2**	**93.9**	**95.1**	**91.9**	**93.9**
$fa \to ar$	4B IT	92.7	84.0	93.8	92.9	91.3	93.2	89.9	94.0	93.2	88.1	91.1	91.2	91.1	91.1	88.4	90.0
	12B IT	96.5	87.1	95.7	95.5	94.8	95.4	92.8	94.5	95.8	93.4	94.1	94.2	94.3	95.0	92.9	94.0
	HexaBee-4B	**98.3**	**91.7**	**98.2**	**97.5**	**96.5**	**97.9**	**95.0**	**97.9**	**97.0**	**95.1**	**95.8**	**95.3**	**96.2**	**95.8**	**94.2**	**94.4**
$es \to ar$	4B IT	90.8	81.0	89.6	90.4	88.7	87.9	86.9	88.0	89.8	83.1	87.1	88.8	89.4	88.2	86.9	90.5
	12B IT	95.5	87.6	**96.1**	**95.3**	94.7	**95.3**	**92.0**	93.5	**95.7**	**93.6**	**92.9**	92.8	**94.5**	93.8	**91.4**	**94.5**
	HexaBee-4B	**95.9**	**89.8**	95.8	**95.3**	**94.8**	95.1	91.3	**95.0**	94.2	91.0	**92.9**	**93.3**	92.7	**94.0**	90.5	94.3

Table 18: Every judge dimension for translations into Arabic, ten shared items per direction. Sem.: semantic accuracy; Term.: terminology; Flu.: fluency; Coh.: cohesion; Ctx.: contextual appropriateness; Lang.: target-language adherence; Fmt.: formatting. Scores are model judgments on 0–100 scales, not validated accuracy percentages. Bold/underline mark first/second within each direction and view; ties at displayed precision share rank.

		Reference-assisted judge								Reference-free judge							
Dir.	**System**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**
$ar \to zh$	4B IT	75.8	72.4	74.9	73.7	74.5	75.1	72.1	74.5	76.7	72.9	75.3	74.4	74.8	75.3	72.4	74.0
	12B IT	95.6	88.5	94.1	93.8	94.7	93.6	90.6	91.5	95.1	91.8	94.4	92.4	93.3	92.4	90.8	91.5
	HEXABEE-4B	**95.9**	**92.1**	**96.5**	**95.3**	**95.6**	**96.3**	**92.5**	**96.4**	**96.0**	**94.0**	**95.4**	**94.1**	**94.5**	**95.7**	**92.5**	**95.0**
$en \to zh$	4B IT	84.8	80.5	80.7	81.1	83.1	82.5	80.0	92.5	86.0	81.7	90.9	93.5	84.2	82.7	81.2	94.5
	12B IT	95.4	**90.4**	91.0	92.4	93.5	92.3	88.8	90.5	96.0	**93.7**	93.2	93.4	94.3	93.9	92.2	93.5
	HEXABEE-4B	**96.3**	90.0	**95.7**	**94.1**	**96.0**	**96.2**	**91.7**	**96.0**	**96.2**	92.5	**95.2**	**94.0**	**94.7**	**95.5**	**92.4**	**95.0**
$he \to zh$	4B IT	80.6	69.1	80.5	80.4	80.4	82.3	77.9	81.5	89.5	79.4	92.8	91.0	90.1	82.7	89.2	93.5
	12B IT	96.0	**91.9**	94.4	94.3	94.9	93.7	91.2	93.0	**96.1**	**94.0**	93.6	93.5	**94.4**	94.1	92.9	94.7
	HEXABEE-4B	**97.1**	91.3	**96.5**	**95.9**	**95.8**	**96.4**	**94.1**	**97.0**	**96.1**	**94.0**	**95.1**	**94.5**	94.2	**95.5**	**93.4**	**95.0**
$fa \to zh$	4B IT	76.5	70.7	75.2	75.9	75.6	75.7	72.0	75.5	85.9	82.3	85.4	85.0	75.6	75.7	74.2	84.8
	12B IT	94.8	90.0	94.9	93.4	93.2	95.3	90.4	94.0	93.8	87.5	93.1	92.5	93.3	94.0	90.8	92.0
	HEXABEE-4B	**97.0**	**92.8**	**96.4**	**95.9**	**95.3**	**96.5**	**94.0**	**96.3**	**96.1**	**93.6**	**96.0**	**94.9**	**95.6**	**95.9**	**94.6**	**94.5**
$es \to zh$	4B IT	72.6	63.3	69.8	71.1	70.6	68.6	66.4	65.0	74.6	68.5	72.0	72.3	72.7	72.5	70.7	70.0
	12B IT	82.7	79.1	82.3	82.4	82.8	82.3	79.8	83.5	86.3	85.5	85.6	85.0	85.8	89.0	84.3	84.0
	HEXABEE-4B	**86.5**	**82.6**	**86.7**	**84.9**	**85.7**	**86.4**	**83.7**	**85.5**	**87.4**	**86.2**	**89.8**	**86.8**	**87.1**	**89.4**	**85.2**	**88.0**

Table 19: Every judge dimension for translations into Chinese, ten shared items per direction. Sem.: semantic accuracy; Term.: terminology; Flu.: fluency; Coh.: cohesion; Ctx.: contextual appropriateness; Lang.: target-language adherence; Fmt.: formatting. Scores are model judgments on 0–100 scales, not validated accuracy percentages. Bold/underline mark first/second within each direction and view; ties at displayed precision share rank.

		Reference-assisted judge								Reference-free judge							
Dir.	**System**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**
$ar \to en$	4B IT	83.3	75.0	85.0	84.0	83.7	85.1	80.6	86.5	86.2	82.8	84.4	84.3	83.7	85.2	81.3	85.5
	12B IT	**95.1**	**89.0**	95.5	94.5	94.5	95.5	91.2	93.0	96.3	94.4	**95.5**	**95.3**	**95.4**	95.7	**93.7**	94.5
	HEXABEE-4B	91.2	87.4	**96.9**	**96.9**	**96.3**	**97.1**	**94.1**	**97.0**	**96.6**	**94.7**	95.2	94.5	**95.4**	**96.2**	93.5	**95.1**
$zh \to en$	4B IT	65.1	61.5	64.1	65.8	64.9	63.8	62.0	75.5	68.0	65.0	75.0	75.4	65.8	66.6	65.3	76.0
	12B IT	**97.9**	**95.3**	95.9	**96.9**	**96.7**	96.1	**94.4**	95.2	**97.1**	**96.4**	94.8	94.8	95.0	96.2	93.1	94.5
	HEXABEE-4B	97.5	93.6	**96.8**	96.4	**96.7**	**96.7**	94.1	**96.0**	**97.1**	**96.4**	**96.8**	**95.4**	**95.9**	**97.1**	**94.6**	**95.8**
$he \to en$	4B IT	**95.6**	**88.8**	96.4	95.7	**95.6**	97.1	92.6	95.0	**93.2**	**90.7**	95.8	**95.0**	**94.5**	**96.2**	93.6	94.0
	12B IT	92.4	86.3	97.1	97.2	95.1	96.8	93.5	96.0	90.3	86.7	**95.9**	94.7	90.5	95.2	93.2	94.3
	HEXABEE-4B	91.4	88.4	**98.4**	**98.2**	92.7	**98.9**	**95.5**	**98.5**	90.9	89.1	94.9	94.8	91.1	95.9	**93.7**	**96.3**
$fa \to en$	4B IT	97.5	94.0	96.9	97.0	97.3	97.0	94.1	94.0	96.5	95.8	94.7	95.9	96.2	95.5	95.0	95.0
	12B IT	**98.4**	**97.5**	97.6	97.7	98.7	97.7	**95.1**	95.8	97.8	96.4	95.6	96.0	96.9	96.7	94.7	94.5
	HEXABEE-4B	97.9	97.3	**99.2**	**97.9**	**98.9**	**99.3**	**95.1**	**98.0**	**98.2**	**97.7**	**96.7**	**96.7**	**97.4**	**97.4**	**96.1**	**95.7**
$es \to en$	4B IT	99.1	96.3	**97.8**	98.3	98.4	**98.2**	95.1	96.7	98.3	98.6	97.3	97.3	97.9	97.5	96.6	96.0
	12B IT	**99.7**	**98.9**	97.0	**99.3**	**99.6**	97.8	**95.5**	97.9	**99.0**	**99.0**	97.3	97.6	**98.7**	97.3	**96.8**	96.3
	HEXABEE-4B	96.2	95.0	97.5	97.4	96.6	97.6	93.6	**98.0**	97.3	96.4	**97.5**	**97.7**	97.1	**97.6**	96.6	**97.5**

Table 20: Every judge dimension for translations into English, ten shared items per direction. Sem.: semantic accuracy; Term.: terminology; Flu.: fluency; Coh.: cohesion; Ctx.: contextual appropriateness; Lang.: target-language adherence; Fmt.: formatting. Scores are model judgments on 0–100 scales, not validated accuracy percentages. Bold/underline mark first/second within each direction and view; ties at displayed precision share rank.

		Reference-assisted judge								Reference-free judge							
Dir.	**System**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**
$ar \to he$	4B IT	72.5	64.0	72.0	72.5	72.0	72.0	68.5	83.0	69.4	62.5	69.5	69.9	69.2	70.0	66.6	69.5
	12B IT	96.0	86.3	94.5	94.4	95.0	95.0	90.0	94.5	96.3	92.9	95.2	93.9	94.1	94.5	92.3	92.1
	HEXABEE-4B	**96.3**	**88.5**	**95.6**	**96.2**	**96.2**	**95.6**	**93.6**	**97.5**	**97.6**	**94.8**	**95.8**	**94.9**	**96.5**	**95.9**	**93.8**	**96.5**
$zh \to he$	4B IT	89.0	80.8	87.7	88.7	88.7	87.4	86.0	88.0	90.4	84.8	87.5	87.9	89.0	89.1	86.5	90.0
	12B IT	93.3	86.8	92.9	91.8	92.3	92.9	90.1	91.2	92.7	87.2	89.0	91.2	88.7	89.5	86.9	91.0
	HEXABEE-4B	**97.4**	**90.5**	**95.9**	**95.6**	**96.0**	**96.5**	**93.3**	**97.3**	**96.4**	**94.8**	**95.0**	**95.3**	**95.1**	**95.6**	**93.2**	**96.0**
$en \to he$	4B IT	72.0	62.5	79.5	80.0	74.5	78.5	72.5	85.5	82.2	69.5	79.5	81.7	80.0	80.0	78.5	87.0
	12B IT	96.6	88.1	94.8	94.8	95.9	93.8	92.1	93.0	95.0	91.1	91.6	**93.2**	93.3	92.4	90.9	93.5
	HEXABEE-4B	**96.9**	**91.2**	**95.6**	**95.7**	**96.6**	**95.5**	**92.4**	**97.0**	**95.6**	**92.5**	**92.8**	93.1	**93.9**	**93.8**	**91.8**	**95.5**
$fa \to he$	4B IT	72.0	61.5	76.3	77.5	71.5	76.3	72.5	90.5	79.0	70.8	78.6	79.2	78.1	80.5	76.1	82.0
	12B IT	95.1	84.0	94.6	94.2	92.3	94.1	**91.3**	**95.5**	94.7	90.2	92.4	91.7	92.6	92.5	90.8	94.0
	HEXABEE-4B	**95.4**	**89.0**	**95.3**	**94.4**	**94.9**	**95.5**	91.1	95.1	**96.6**	**94.9**	**94.6**	**94.7**	**95.9**	**95.7**	**92.5**	**95.6**
$es \to he$	4B IT	88.3	78.7	85.7	87.6	86.1	84.7	83.5	93.0	88.7	82.2	87.2	88.8	88.2	87.6	86.7	93.0
	12B IT	94.0	86.5	94.2	93.1	93.8	94.2	90.4	94.0	94.2	90.0	93.4	92.6	92.3	93.1	90.8	93.0
	HEXABEE-4B	**97.2**	**92.2**	**96.0**	**95.2**	**95.1**	**95.7**	**93.0**	**97.5**	**97.3**	**95.3**	**95.3**	**95.2**	**95.8**	**95.5**	**93.6**	**96.0**

Table 21: Every judge dimension for translations into Hebrew, ten shared items per direction. Sem.: semantic accuracy; Term.: terminology; Flu.: fluency; Coh.: cohesion; Ctx.: contextual appropriateness; Lang.: target-language adherence; Fmt.: formatting. Scores are model judgments on 0–100 scales, not validated accuracy percentages. Bold/underline mark first/second within each direction and view; ties at displayed precision share rank.

		Reference-assisted judge								Reference-free judge							
Dir.	**System**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**
$ar \to fa$	4B IT	77.0	68.8	73.4	74.8	75.5	75.3	71.0	78.0	77.6	74.1	76.0	76.3	76.2	76.6	74.3	75.5
	12B IT	**93.2**	85.4	94.1	93.0	91.1	94.1	89.3	94.5	95.4	91.0	95.0	93.5	94.1	94.8	91.8	94.0
	HEXABEE-4B	92.2	**85.7**	**96.4**	**94.5**	**94.6**	**96.4**	**92.0**	**97.7**	**97.0**	**95.2**	**96.5**	**96.4**	**96.8**	**96.1**	**95.3**	**96.5**
$zh \to fa$	4B IT	51.5	45.0	53.0	51.0	51.5	51.5	49.0	64.5	55.0	49.5	55.0	55.0	55.5	54.0	53.0	56.5
	12B IT	91.4	84.8	93.3	92.4	92.3	93.0	90.0	93.0	94.3	89.8	92.3	93.1	93.6	93.6	90.8	92.5
	HEXABEE-4B	**93.6**	**86.4**	**95.1**	**93.1**	**94.3**	**95.4**	**92.9**	**94.5**	**95.2**	**92.6**	**94.5**	**93.8**	**94.6**	**94.9**	**92.8**	**96.0**
$en \to fa$	4B IT	94.1	85.7	92.6	92.9	91.0	91.4	88.4	92.0	93.2	87.2	88.7	91.6	90.1	89.4	87.3	91.5
	12B IT	95.4	88.0	93.4	93.8	94.0	93.4	90.7	92.5	95.1	92.3	91.5	92.6	92.7	92.8	89.4	93.3
	HEXABEE-4B	**96.7**	**90.3**	**94.4**	**95.2**	**95.2**	**94.3**	**91.8**	**93.5**	**96.3**	**94.5**	**93.4**	**94.0**	**94.1**	**94.1**	**92.1**	**94.0**
$he \to fa$	4B IT	94.8	84.9	92.8	93.3	94.5	93.1	90.2	93.5	94.7	90.9	93.1	93.4	93.4	93.1	91.4	93.5
	12B IT	97.0	89.4	95.5	95.8	96.1	95.6	93.0	94.5	96.5	94.3	94.6	95.5	96.0	94.8	94.4	94.5
	HEXABEE-4B	**97.7**	**90.9**	**97.5**	**96.9**	**96.7**	**97.3**	**94.3**	**96.9**	**98.0**	**96.9**	**96.3**	**95.7**	**96.8**	**96.5**	**94.5**	**95.0**
$es \to fa$	4B IT	82.1	74.4	80.2	81.2	81.4	80.7	80.0	80.5	84.8	79.4	81.9	83.6	82.7	81.9	80.4	83.0
	12B IT	95.5	86.7	93.6	93.7	93.9	92.7	90.8	91.0	95.4	91.8	91.8	**93.5**	94.1	92.1	91.2	94.0
	HEXABEE-4B	**96.5**	**88.2**	**94.3**	**95.4**	**95.6**	**94.7**	**92.9**	**96.0**	**96.2**	**94.4**	**93.5**	93.4	**94.8**	**94.7**	**92.2**	**95.0**

Table 22: Every judge dimension for translations into Persian, ten shared items per direction. Sem.: semantic accuracy; Term.: terminology; Flu.: fluency; Coh.: cohesion; Ctx.: contextual appropriateness; Lang.: target-language adherence; Fmt.: formatting. Scores are model judgments on 0–100 scales, not validated accuracy percentages. Bold/underline mark first/second within each direction and view; ties at displayed precision share rank.

		Reference-assisted judge								Reference-free judge							
Dir.	**System**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**
$ar \to es$	4B IT	91.8	80.7	94.1	92.2	92.3	93.0	89.1	92.5	94.4	87.8	94.3	94.0	93.4	93.5	92.5	95.8
	12B IT	**97.4**	**89.5**	**98.0**	**97.1**	**96.9**	**97.3**	**93.9**	93.0	**98.5**	**96.5**	96.6	**96.7**	**97.5**	96.6	**95.9**	96.0
	HEXABEE-4B	95.0	85.5	96.8	95.8	96.1	96.1	93.8	**96.8**	96.2	91.6	**97.1**	95.7	95.3	**97.1**	94.3	**96.1**
$zh \to es$	4B IT	86.1	78.0	85.0	84.7	85.4	84.9	81.7	83.5	87.6	87.3	85.6	86.4	86.7	86.4	84.6	85.5
	12B IT	96.4	90.6	95.7	95.1	95.5	95.5	92.0	93.5	95.8	93.3	92.9	93.6	93.2	93.3	91.5	93.6
	HEXABEE-4B	**97.5**	**91.4**	**96.6**	**95.2**	**96.8**	**96.3**	**94.3**	**96.5**	**96.5**	**95.7**	**94.8**	**95.1**	**95.1**	**95.9**	**94.0**	**94.8**
$en \to es$	4B IT	96.9	89.8	95.5	95.8	95.7	94.6	92.9	93.5	96.9	95.6	95.3	95.3	96.6	95.8	94.4	95.0
	12B IT	**98.2**	90.7	**97.3**	97.2	**97.2**	96.4	**94.6**	94.5	**98.3**	**97.3**	95.5	**97.0**	**97.5**	96.2	95.2	**97.5**
	HEXABEE-4B	97.3	**91.6**	96.9	**97.3**	97.1	**96.8**	94.4	**97.5**	97.5	96.4	**96.9**	96.2	96.6	**97.0**	**95.6**	96.2
$he \to es$	4B IT	80.6	75.0	79.7	80.1	79.6	79.3	76.9	95.5	78.2	76.8	77.7	77.7	77.3	78.2	75.9	80.5
	12B IT	**99.2**	93.6	**98.9**	**98.5**	98.6	98.3	**95.7**	95.7	87.8	87.2	86.8	86.8	87.5	86.7	86.1	85.5
	HEXABEE-4B	99.1	**97.5**	98.3	98.2	**99.0**	**98.9**	94.6	**99.0**	**98.1**	**98.0**	**98.2**	**97.6**	**98.1**	**98.4**	**95.8**	**95.8**
$fa \to es$	4B IT	84.6	81.2	94.5	93.2	88.7	93.9	87.3	92.5	92.5	90.9	94.8	92.9	92.4	94.9	**91.6**	91.0
	12B IT	97.5	94.0	96.3	96.8	95.5	96.5	92.5	94.0	95.0	92.6	**95.3**	93.3	92.4	**95.5**	91.5	91.0
	HEXABEE-4B	**98.1**	**94.2**	**96.4**	**97.0**	**97.1**	**97.7**	**93.3**	**96.1**	**96.2**	**93.8**	94.0	**94.1**	**93.1**	95.1	**91.6**	**94.5**

Table 23: Every judge dimension for translations into Spanish, ten shared items per direction. Sem.: semantic accuracy; Term.: terminology; Flu.: fluency; Coh.: cohesion; Ctx.: contextual appropriateness; Lang.: target-language adherence; Fmt.: formatting. Scores are model judgments on 0–100 scales, not validated accuracy percentages. Bold/underline mark first/second within each direction and view; ties at displayed precision share rank.

		Reference-assisted judge								Reference-free judge							
Target	**System**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**	**Sem.**	**Term.**	**Flu.**	**Coh.**	**Ctx.**	**Lang.**	**Style**	**Fmt.**
ar	4B IT	85.1	77.1	85.5	85.3	84.3	84.9	81.6	87.0	84.4	78.3	83.6	84.0	83.2	84.0	81.5	85.5
	12B IT	92.8	84.3	94.5	93.9	91.5	92.0	89.0	92.9	94.3	90.6	92.2	92.3	92.9	90.9	91.2	93.1
	HEXABEE-4B	**95.7**	**89.8**	**96.6**	**95.4**	**94.8**	**96.1**	**92.6**	**96.4**	**95.5**	**93.4**	**94.4**	**94.3**	**94.5**	**95.0**	**92.3**	**94.6**
zh	4B IT	78.1	71.2	76.2	76.4	76.8	76.8	73.7	77.8	82.5	77.0	83.3	83.2	79.5	77.8	77.5	83.4
	12B IT	92.9	88.0	91.3	91.3	91.8	91.4	88.2	90.5	93.5	90.5	92.0	91.4	92.2	92.7	90.2	91.1
	HEXABEE-4B	**94.6**	**89.8**	**94.4**	**93.2**	**93.7**	**94.4**	**91.2**	**94.2**	**94.4**	**92.1**	**94.3**	**92.9**	**93.2**	**94.4**	**91.6**	**93.5**
en	4B IT	88.1	83.1	88.0	88.2	88.0	88.2	84.9	89.5	88.4	86.6	89.4	89.6	87.6	88.2	86.4	89.3
	12B IT	**96.7**	**93.4**	96.6	97.1	**96.9**	96.8	93.9	95.6	**96.1**	94.6	95.8	95.7	95.3	96.2	94.3	94.8
	HEXABEE-4B	94.8	92.3	**97.8**	**97.4**	96.2	**97.9**	**94.5**	**97.5**	96.0	**94.9**	**96.2**	**95.8**	**95.4**	**96.8**	**94.9**	**96.1**
he	4B IT	78.8	69.5	80.2	81.3	78.6	79.8	76.6	88.0	81.9	74.0	80.5	81.5	80.9	81.4	78.9	84.3
	12B IT	95.0	86.3	94.2	93.7	93.9	94.0	90.8	93.6	94.6	90.3	92.3	92.5	92.2	92.4	90.3	92.7
	HEXABEE-4B	**96.6**	**90.3**	**95.7**	**95.4**	**95.8**	**95.8**	**92.7**	**96.9**	**96.7**	**94.5**	**94.7**	**94.6**	**95.4**	**95.3**	**93.0**	**95.9**
fa	4B IT	79.9	71.8	78.4	78.6	78.8	78.4	75.7	81.7	81.1	76.2	78.9	80.0	79.6	79.0	77.3	80.0
	12B IT	94.5	86.9	94.0	93.7	93.5	93.8	90.8	93.1	95.3	91.8	93.0	93.6	94.1	93.6	91.5	93.7
	HEXABEE-4B	**95.3**	**88.3**	**95.5**	**95.0**	**95.3**	**95.6**	**92.8**	**95.7**	**96.5**	**94.7**	**94.8**	**94.7**	**95.4**	**95.3**	**93.4**	**95.3**
es	4B IT	88.0	80.9	89.8	89.2	88.3	89.1	85.6	91.5	89.9	87.7	89.5	89.3	89.3	89.8	87.8	89.6
	12B IT	**97.7**	91.7	**97.2**	**96.9**	96.7	96.8	93.7	94.1	95.1	93.4	93.4	93.5	93.6	93.7	92.0	92.7
	HEXABEE-4B	97.4	**92.0**	97.0	96.7	**97.2**	**97.2**	**94.1**	**97.2**	**96.9**	**95.1**	**96.2**	**95.7**	**95.6**	**96.7**	**94.3**	**95.5**

Table 24: All eight judge dimensions in both views, aggregated by target language, with 50 shared examples per target. All directions receive equal weight.

Split	System	N	Src.	Cand.	Ref.
Full	4B IT	36,572	7757	8240	7789
Full	12B IT	36,572	7757	8334	7789
Full	HEXABEE-4B	36,572	7757	7743	7789
Judge	4B IT	300	51	52	52
Judge	12B IT	300	51	53	52
Judge	HEXABEE-4B	300	51	51	52

Table 25: Texts exceeding the 512-token COMET encoder limit. Src., Cand., and Ref. denote source, candidate, and reference. Counts can overlap within an example.

Metric settings. SacreBLEU 2.5.1: one reference, mixed-case BLEU, exponential smoothing, no effective order, `zh` tokenization for Chinese targets and `13a` otherwise. chrF and chrF++ use character order six, $\beta = 2$, and respectively zero/two word orders, excluding character whitespace. TER uses lowercase Tercom tokenization with normalization, punctuation retained, and Asian support. The compiled TER beam kernel preserves the reference operation tie order, beam, shift limits, and prefix cache; exact edit/length totals match SacreBLEU on all 900 judge-subset outputs, alongside 2,000 exact score/trace cache checks and 250 candidate-order checks. The compiled complete search additionally matches 500 randomized cases; the reference tokenizer and search semantics are retained.

COMET. We use `Unbabel/wmt22-comet-da` with unbabel-comet 2.2.7, official checkpoint revision `2760a223ac95`, and verified SHA-256. Inference uses float32 (batch 32 for full-test scoring, 16 for the judge subset). Values in tables are raw scores multiplied by 100 and are not accuracy percentages. The encoder truncates each source, candidate, and reference separately at 512 tokens, so long-document suffixes are not assessed by this metric.

G Evaluation integrity and reproducibility

The parser used truth-value fallback for the score field, which turned some explicit zero scores into nulls in intermediate records. We reparse the original judge JSON and recover 11 reference-view and 13 reference-free zeros for the 4B baseline, plus one reference-free zero for the 12B baseline. All 300 items then have valid scores in both views. Dropping these entries would inflate baseline means. The reparsed means agree with the final flattened score files.

Content identity is the SHA-256 hash of a canonical JSON list containing source language, target language, source text, and reference. We require equality of identity sets before comparing full-test files and before pairing judge scores. File row indices are not stable identities: asynchronous processing and sampling changed order across models. For judge recovery, an explicit JSON field `score: 0` remains zero rather than being treated as a missing value.

Our audit scripts are read-only with respect to the original training and evaluation artifacts. They preserve null/empty outputs, pin metric settings, save directional sufficient statistics where applicable, and produce content-hash IDs without copying raw evaluation text into the paper package. The local evidence directory records the actual training/merge paths and sanitized snapshots for author verification; those private paths are omitted from this release. The paper reports a retrospective analysis of existing translation outputs and newly computed metrics, not new training runs or newly collected human labels.