VISTA-BENCH: BENCHMARKING MULTILINGUAL IMAGE TRANSLATION WITH IMAGE-SPECIFIC RUBRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image translation is a fundamental capability of multimodal models for multilingual applications, requiring visual understanding and meaning preservation across languages. However, existing benchmarks have limited language coverage and often lack explicit image-specific evaluation criteria, making it difficult to comprehensively assess this capability. To systematically evaluate this capability, we introduce VISTA-Bench, covering 22 languages and 10 domains, and develop an image-specific rubric evaluation protocol. The benchmark combines sampling for language and scenario coverage with model-assisted, human-verified annotations that group related text into coherent semantic units and provide multilingual reference translations. The rubrics specify essential content, semantic relations, and acceptable translation variants, yielding separate output-based scores for translation quality and the preservation of visual and knowledge-dependent information. We conduct extensive evaluations of 16 mainstream models, including 12 multimodal models and four text-input models, and provide systematic analyses across languages, domains, and evaluation dimensions. We will release all code and data.

1 INTRODUCTION

Image translation requires coordinating visual information acquisition, semantic interpretation, and target-language generation to preserve meaning across languages (Zhu et al., 2023). For multimodal large language models, this goes beyond recognizing isolated text: models must interpret relationships among text regions and use visual context to resolve ambiguity where necessary (Qian et al., 2024). We study image-in, translated-text-out evaluation; text localization, layout reconstruction, background preservation, and translated-image quality are outside its scope. A suitable benchmark must therefore combine diverse language and scene coverage with evaluation that accommodates valid translation variation while detecting omissions or mistranslations of essential information and semantic relationships.

Existing benchmarks have advanced visually situated translation, multilingual coverage, and position-aware evaluation (Salesky et al., 2024; Li et al., 2025; Zhuang et al., 2025). These benchmarks employ automatic translation metrics such as BLEU (Papineni et al., 2002), chrF (Popović, 2015), and COMET (Rei et al., 2020). Recent work also incorporates multimodal model-based judging (Li et al., 2026). Nevertheless, source- and target-language coverage remains uneven, and individual benchmarks emphasize different subsets of real-world scenarios. On the evaluation side, aggregate similarity scores and general quality judgments do not necessarily reveal whether a translation satisfies the specific requirements of an image. Visual-dependence annotations and localization metrics provide useful complementary information, but they serve different purposes from explicitly specifying which meanings, relations, and acceptable interpretations should be preserved.

To address these needs, we introduce VISTA-Bench, comprising 2,228 images across 22 languages, 10 major domains, and 100+ fine-grained scenarios. Its construction jointly considers source-language and scenario coverage, using sampling across language–scenario groups to obtain a compact evaluation set. Model-assisted annotation and human verification establish the source text, semantic blocks, and multilingual reference translations (Appendix A). In particular, annotators merge fragments that jointly express a semantic unit while retaining the separation of unrelated con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tent. This process organizes the benchmark around meaningful source content rather than treating individual text detections as independent translation units.

Prior work explores fine-grained MQM error annotation with LLMs (Kocmi & Federmann, 2023) and case-specific MQM rubrics (Xu et al., 2026). Alongside the data, we develop an image-specific rubric evaluation protocol. Each image receives a separate rubric covering visual understanding, translation quality, and knowledge-dependent interpretation where applicable. Shared criteria are supplemented with target-language-specific requirements when needed. The criteria specify required meaning, acceptable expressions, and disallowed errors, allowing legitimate variation while making content-specific constraints explicit. A judge model assesses the candidate translation against these criteria and provides item-level decisions with supporting evidence. Fixed rules aggregate the decisions by criterion importance within each dimension, yielding separate V, T, and K scores. Our contribution thus combines coverage-oriented benchmark construction with an evaluation design that connects translation scores to identifiable requirements in individual images.

The remainder of this paper is organized as follows. After reviewing related work, we describe the construction of VISTA-Bench and its image-specific rubric evaluation protocol. We then benchmark 16 mainstream models, comprising 12 multimodal models and four text-input models, across 22 languages and 10 domains. We analyze translation quality alongside visual understanding and knowledge use, examine variation across languages and domains, and assess dataset-size sensitivity and agreement between automatic judges and human annotators.

2 Related Work

2.1 Text-Image Machine Translation

Text-image machine translation (TIMT) translates source-language text embedded in images into a target language (Lan et al., 2023). Existing approaches address both cross-modal alignment and contextual interpretation: PEIT uses pretrained models to bridge the vision–text modality gap in end-to-end translation (Zhu et al., 2023), while AnyTrans incorporates textual and visual context when translating fragmented text (Qian et al., 2024). These complementary directions connect text perception with the preservation of meaning and relationships among text regions.

2.2 Multilingual Image Translation Benchmarks

Among existing multilingual image translation benchmarks (Salesky et al., 2024; Li et al., 2025; Tian et al., 2025a; Li et al., 2026), Vistra and PRIM both use only English source images, with translations into four and five target languages, respectively. Vistra focuses on visual context in natural images, whereas PRIM primarily covers images containing a single line of text (Salesky et al., 2024; Tian et al., 2025a). MIT-10M contains source images in eight languages, mainly drawn from web pages and product displays (Li et al., 2025). MMTIT-Bench includes images in fourteen source languages other than English and Chinese, but provides translations only into Chinese and English, covering documents, web pages, and natural scenes (Li et al., 2026).

Across these benchmarks, source–target coverage and scenario emphasis remain uneven. Evaluation includes BLEU, chrF, and COMET (Salesky et al., 2024), as well as multimodal judging with shared quality dimensions (Li et al., 2026). In text translation, GEMBA-MQM identifies fine-grained error spans (Kocmi & Federmann, 2023), while Rubric-as-Experts develops case-specific MQM rubrics (Xu et al., 2026). Building on these lines of work, VISTA-Bench combines multilingual, multi-scenario coverage with image-specific rubrics that specify required meanings, semantic relations, and acceptable translation variants.

3 Benchmark Construction and Evaluation

VISTA-Bench comprises 2,228 images across 22 languages, 10 major domains, and 100+ fine-grained scenarios. The annotation inventory distinguishes two Chinese scripts, giving 23 source/target codes and 49,016 reference slots; after merging script variants and excluding same-language tasks, there are 46,788 nominal image–target-language pairs. Actual scoring coverage is

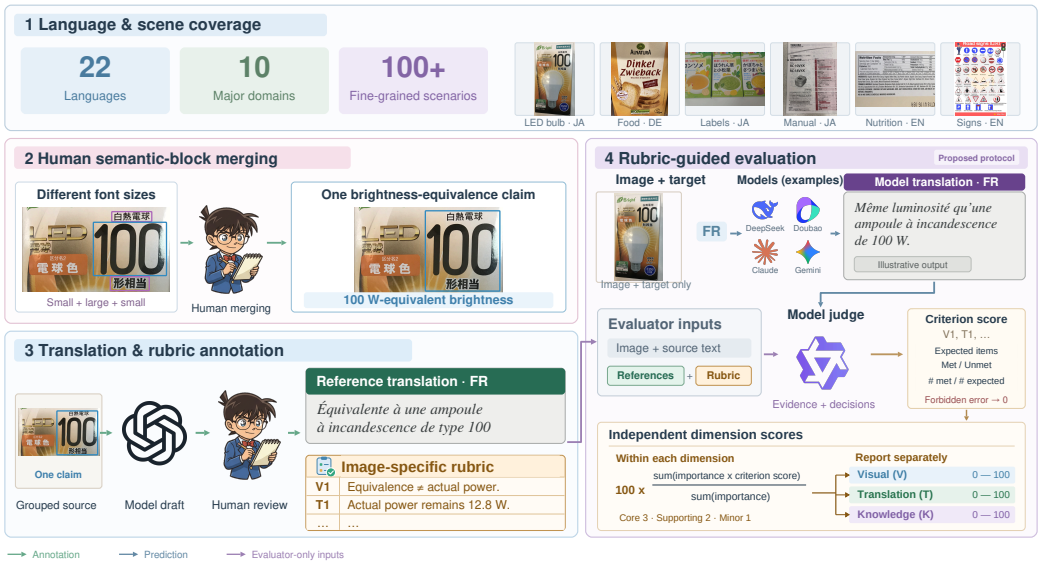


Figure 1: Overview of VISTA-Bench. Stages 1–3 construct language–scenario coverage, human-reviewed semantic blocks, and multilingual references with image-specific rubrics. Stage 4 separates the model's image–target input from evaluator-only annotations, then converts evidence-based judgments into separate importance-weighted V, T, and K scores.

smaller and dimension-specific. Figure 1 summarizes construction and evaluation. Table 1 compares coverage and evaluation protocols across related benchmarks (Salesky et al., 2024; Li et al., 2026; Tian et al., 2025a;b; Zhuang et al., 2025); their text, rendered-image, and localization outputs are not identical.

Table 1: Comparison of image translation benchmarks and evaluation protocols.

Dataset	Test Size	Languages (source → target)	# Scenarios	Evaluation
Vistra	772	En→De/Es/Ru/Zh	No scenario breakdown	BLEU, chrF, COMET
MMTIT-Bench	1,400	14 languages → En/Zh	3 (broad types)	COMET + VLLM judge
PRIM	340	En→De/Fr/Cs/Ru/Ro	No scenario breakdown	BLEU, COMET
IIMT30K	2,740	En↔De	No scenario breakdown	BLEU, COMET
PATIMT-Bench	1,200	En↔Zh	10 (6 eval. groups)	BLEU, COMET
VISTA-Bench (Ours)	2,228	**22 languages → 22 languages**	**100+ (10 domains)**	**Image-specific rubric judge**

3.1 Language and Scenario Coverage

Images from existing collections, local data, and web retrieval undergo source-specific automatic filtering, exact deduplication, and sensitive-information handling (Figure 1, stage 1). We then assign language labels and classify images using a two-level taxonomy of major domains and fine-grained usage scenarios, producing a candidate pool of 10,920 images. For example, printed menus are grouped under dining, road signs under transportation, and textbook content under publications.

To cover diverse use cases within the ten major domains, we use their fine-grained scenarios as sampling units. This allows each domain to be represented by images from different subscenarios while keeping the evaluation set compact. Let $G_1, \ldots, G_K$ denote the nonempty groups, each containing images with the same source language and fine-grained scenario. We form an initial subset by drawing a random sample from each group:

$$x_k \sim \mathrm{Uniform}(G_k), \quad k = 1, \ldots, K, \qquad \mathcal{S}_0 = \{x_k\}_{k=1}^{K}. \tag{1}$$

We then expand this initial selection by sampling additional images from the fine-grained scenario groups in turn within each language, until the target sample size is reached. All selected images undergo human review.

3.2 Human Semantic-Block Merging

Accurate image translation relies on understanding how text regions jointly convey meaning, rather than interpreting each fragment in isolation. We use `gpt5.6 sol` for OCR and layout-aware semantic grouping (Figure 1, stage 2). Annotators correct text, omissions, reading order, and block boundaries against the images, merging fragments that express a shared meaning while keeping unrelated content separate.

"Any revision" denotes the proportion of images with at least one text or structural change, counting each image only once. Across the ten languages in Figure 2, the mean per-language rate of text or structural revision is 42.7%. Rates are computed on images with comparable machine and human bounding boxes, excluding text-only annotations. Structural revisions denote unmatched blocks under one-to-one matching at IoU ≥ 0.5, excluding coordinate-only edits to matched blocks.

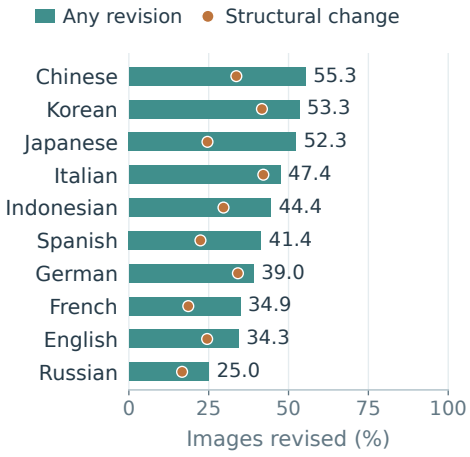


Figure 2: Human revision rates for ten selected widely used languages.

3.3 Reference and Rubric Annotation

We use `gpt5.6 sol` to draft both reference translations and rubrics, with human verification at each stage (Figure 1, stage 3). Translations into the target languages are drafted from the verified source text, with web search to resolve uncertain names and specialized terms. Human annotators check and revise these drafts against the images. The verified source blocks and references then guide rubric drafting, followed by human review and refinement. Appendix B provides a complete rubric example, explains its use, and documents the drafting and judge prompt templates.

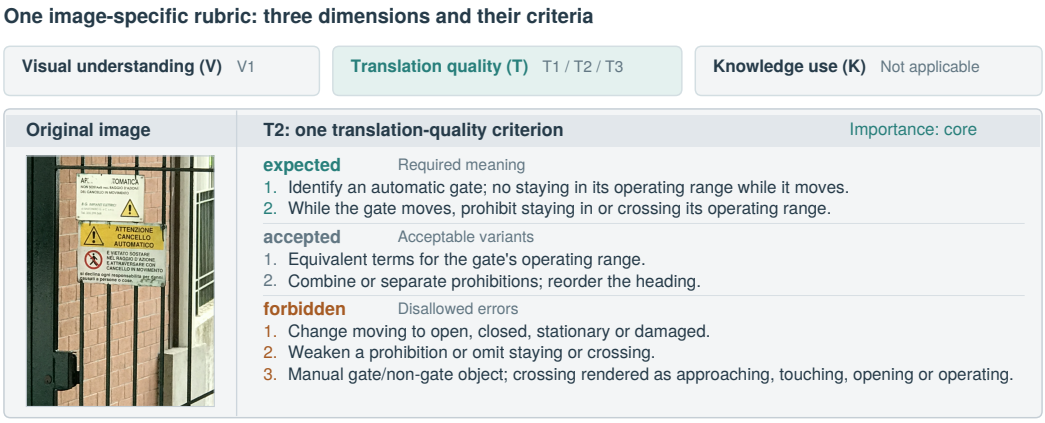


Figure 3: Rubric structure for an automatic-gate notice. Only T2 is expanded, with condensed wording. The knowledge dimension is inapplicable to this image.

To account for lexical ambiguity and variation in valid translations, we construct a separate rubric for each image to complement its verified reference translations. Each rubric defines three output-based

diagnostics: V checks preservation of text coverage, grouping, and visual associations; T checks semantic fidelity and natural target-language expression; and K checks knowledge-dependent interpretations. Knowledge criteria apply only where needed and do not reward additional facts. These scores describe which requirements the translation satisfies, not whether the model actually relied on a particular visual cue or knowledge source.

For image i and target language ℓ, the evaluation criteria are

$$\mathcal{R}_{i,\ell} = \mathcal{R}_i \cup \Delta\mathcal{R}_{i,\ell}, \tag{2}$$

where $\mathcal{R}_i$ contains the image's shared criteria and $\Delta\mathcal{R}_{i,\ell}$ contains target-language-specific additions, with $\Delta\mathcal{R}_{i,\ell} = \varnothing$ when none are needed. Each criterion specifies required meaning (`expected`), acceptable expressions (`accepted`), and disallowed errors (`forbidden`). Its importance label (core, supporting, or minor) determines its weight in score aggregation (Section 3.4).

3.4 Evaluation Protocol

Once the rubrics are established, the evaluated model generates a free-form translation $\hat{y}_{i\ell} = M(I_i, \ell)$ given only the image I_i and target language ℓ, without access to the annotations. A judge model then assesses this output against the corresponding rubric and verified source blocks, returning item-level judgments and supporting evidence rather than an overall rating. The protocol uses fixed rules to compute scores from these judgments, restricting score assignment while leaving semantic judgments subject to bias.

The judge marks each expected item as met only when its requirement is fully satisfied, and as unmet otherwise. For a given image–target pair, let $n_j > 0$ denote the total number of expected items in criterion j and n_j^{met} the number met. The score is the proportion of items met unless a forbidden error occurs. Writing $f_j = 1$ when any listed forbidden error is found and $f_j = 0$ otherwise, we obtain

$$q_j = \begin{cases} 0, & f_j = 1, \\ \dfrac{n_j^{\text{met}}}{n_j}, & f_j = 0. \end{cases} \tag{3}$$

Only expected items enter the denominator. Accepted variants can satisfy these items but earn no additional credit. A forbidden error sets only the corresponding criterion score to zero. Otherwise, meeting two of three expected items yields $2/3$.

For each image–target pair, we compute a separate importance-weighted score for each applicable dimension,

$$s_{i\ell,d} = 100\frac{\sum_{j\in\mathcal{J}_{i\ell,d}} w_j q_j}{\sum_{j\in\mathcal{J}_{i\ell,d}} w_j}, \qquad d \in \{V, T, K\}, \quad \mathcal{J}_{i\ell,d} \neq \varnothing. \tag{4}$$

Here, $\mathcal{J}_{i\ell,d}$ contains the applicable criteria in dimension d, and $s_{i\ell,d}$ is its score on a 0–100 scale. The importance weights $w_j = 3, 2, 1$ correspond to core, supporting, and minor criteria, respectively, and apply only within a dimension. A dimension with no applicable criteria is marked as not applicable rather than assigned zero. We report V, T, and K separately without combining them into a composite score. T is the primary translation measure, while V and K provide complementary diagnostics.

4 Experiments

4.1 Models

We report results for 16 models: 12 multimodal models and four text-input baselines. The multimodal group includes gpt-6-astra[1], doubao-seed-2-1-pro-260915, gemini-3.8-flash[2], qwen3.8-max-

[1] https://developers.openai.com/api/docs/models/gpt-6-astra

[2] https://deepmind.google/models/model-cards/gemini-3-8-flash

0902[3], muse-spark-1.3[4], claude-sonnet-5[5], qwen3.8-27b-fp8[6], deepseek-flash, gpt-5.6-luna[7], glm-5.3-flash[8], glm-4.1v-9b, and gemma-4-e4b (Gemma Team, 2026). These models receive the original image and a target-language instruction, following Section 3.4.

The four text-input baselines reported in Table 2 are aya-23-8B-text (Aryabumi et al., 2024), gemma-3-4b-it-text (Gemma Team, 2025), HY-MT2-30B-A3B-text (Zheng et al., 2026), and Qwen3-4B-text (Qwen Team, 2025). The `-text` suffix identifies text-input runs; no results for other text-input configurations are claimed here. For the reported text-input results, source text is organized using human-annotated layouts, and the models do not receive the original image.

4.2 Evaluation Scores

We report T, V, and K independently on a 0–100 scale, with T as the primary measure and V and K as complementary diagnostics. For image-input models, retained criterion scores are importance-weighted within each dimension and direction, and global scores average directions equally. Domain T applies the same procedure within each domain over common valid directions. K includes only applicable criteria. The Overall column in Table 2 denotes global T, not a composite of the three dimensions.

4.3 Domain-Level Performance

Table 2 reports global and domain-level T for image-input models, grouped by weight availability, and four text-input baselines. Rows are ordered by global T from low to high within each group; tinted cells mark the best value in each group and column.

Table 2: Translation-quality (T) results. Overall denotes global T, not the composite score. Global T averages directions for image-input models and tasks for text-input models. Tinted cells mark the highest score within each group and column.

Model	Overall	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Image-input: closed-weight models											
GPT-5.6 Luna	55.88	51.00	60.03	53.44	54.38	61.84	60.94	58.92	58.59	58.35	63.55
Claude Sonnet 5	69.43	64.71	71.78	65.71	71.14	74.18	77.94	67.08	71.38	78.13	77.82
Muse Spark 1.3	71.56	64.09	71.98	70.13	71.99	77.43	77.59	73.70	76.86	72.75	73.95
Qwen3.8 Max	72.67	66.00	71.94	72.57	72.89	76.20	81.07	74.67	75.57	76.43	75.40
Gemini 3.8 Flash	74.20	69.10	75.62	74.57	76.22	79.52	79.12	72.64	76.34	78.40	78.17
Doubao Seed 2.1 Pro	75.12	68.43	73.48	75.66	74.24	77.60	81.89	77.48	78.38	80.20	78.87
GPT-6 Astra	77.63	73.78	76.03	76.11	77.02	76.98	83.73	81.63	80.19	79.70	79.17
Image-input: open-weight models											
Gemma-4 E4B	19.54	20.55	12.46	17.12	23.60	23.85	27.74	23.53	17.65	15.83	30.80
GLM-4.1V 9B	29.96	30.39	25.36	27.78	31.06	36.42	33.67	31.99	32.86	38.78	35.71
GLM-5.3 Flash	55.33	52.41	55.44	53.89	54.47	63.23	60.37	55.24	60.74	64.68	60.62
DeepSeek Flash	61.04	52.41	62.95	57.80	63.44	69.86	69.91	62.51	65.88	68.92	68.15
Qwen3.8 27B	61.11	56.79	55.63	60.18	63.22	68.80	67.59	63.18	67.32	65.60	68.52
Text-input models											
Aya-23 8B	47.43	48.97	41.83	52.95	51.41	54.14	49.67	46.79	46.73	40.91	48.96
Gemma-3 4B IT	55.99	56.22	52.28	60.13	57.92	63.03	58.72	54.22	55.29	50.11	58.52
Qwen3 4B	64.34	65.44	54.00	68.55	66.27	72.38	67.10	61.73	65.56	59.56	67.60
HY-MT2 30B-A3B	72.82	72.75	71.05	78.02	75.65	81.67	75.42	68.12	75.22	64.56	76.15

Among the image-input models, GPT-6 Astra achieves the highest global T (77.63) and ranks first in eight of the ten domains. The exceptions are Healthcare, led by Gemini 3.8 Flash (79.52), and

[3] https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max
[4] https://dev.meta.ai/
[5] https://www.anthropic.com/news/claude-sonnet-5
[6] https://huggingface.co/Qwen/Qwen3.8-27B-FP8
[7] https://developers.openai.com/api/docs/models/gpt-5.6-luna
[8] https://huggingface.co/zai-org/GLM-5.3-Flash

Finance, led by Doubao Seed 2.1 Pro (80.20). Within the open-weight image-input group, Qwen3.8 27B has the highest global T (61.11), while DeepSeek Flash leads in Dining, Tourism, Healthcare, Safety, and Finance. These changes in ranking show why domain-level scores complement the global average. Appendix C.1 reports domain T by translation target language, averaging over source languages.

With human-organized source text, HY-MT2 30B-A3B achieves a global T of 72.82, compared with 77.63 for GPT-6 Astra. It exceeds the best image-input scores in Products (78.02 vs. 76.11) and Healthcare (81.67 vs. 79.52), but trails in Finance (64.56 vs. 80.20). The supplied semantic blocks let the text-input model bypass image recognition and layout recovery, whereas image-input models must recover both the text and its organization from pixels. This contrast highlights text grouping and reading order as additional requirements of image translation.

4.4 Language-Level Performance

Figure 4 reports source-language T profiles for six representative models.



Figure 4: Source-language translation profiles for six multimodal models. Each point is the mean T for one source language over its target-language directions. Dashed lines mark the per-model average across 22 languages; teal and orange denote scores above and below that average.

The profiles expose both common bottlenecks and model-specific strengths. Burmese is the lowest-scoring source language for all six models, ranging from 22.90 for GLM-5.3 Flash to 51.22 for GPT-6 Astra. Dutch is strongest for Gemini 3.8 Flash, Claude Sonnet 5, Qwen3.8 27B, and GLM-5.3 Flash, whereas Czech is strongest for GPT-6 Astra and DeepSeek Flash. GPT-6 Astra obtains the highest score among these models on 19 of the 22 languages. The three China-developed models shown all exceed their own language averages on Chinese: Qwen3.8 27B scores 64.79 versus 60.96, DeepSeek Flash 61.78 versus 61.06, and GLM-5.3 Flash 58.82 versus 55.21. By contrast, GPT-6

Astra, Gemini 3.8 Flash, and Claude Sonnet 5 score 2.20–3.95 points below their respective averages on Chinese. The shared difficulty on Burmese and differing language strengths show the value of broad source-language coverage: it exposes both common weaknesses and capabilities concentrated in particular languages. These profiles identify where multilingual image translation needs further improvement. Complete source-to-target T matrices for all 12 models are provided in Appendix C.2.

4.5 Dimension-Level Analysis

Figure 5 reports V and K alongside T for all 12 models, using the same T-based row order in both panels. All three metrics use benchmark-wide direction averages; K retains its dimension-specific applicability.

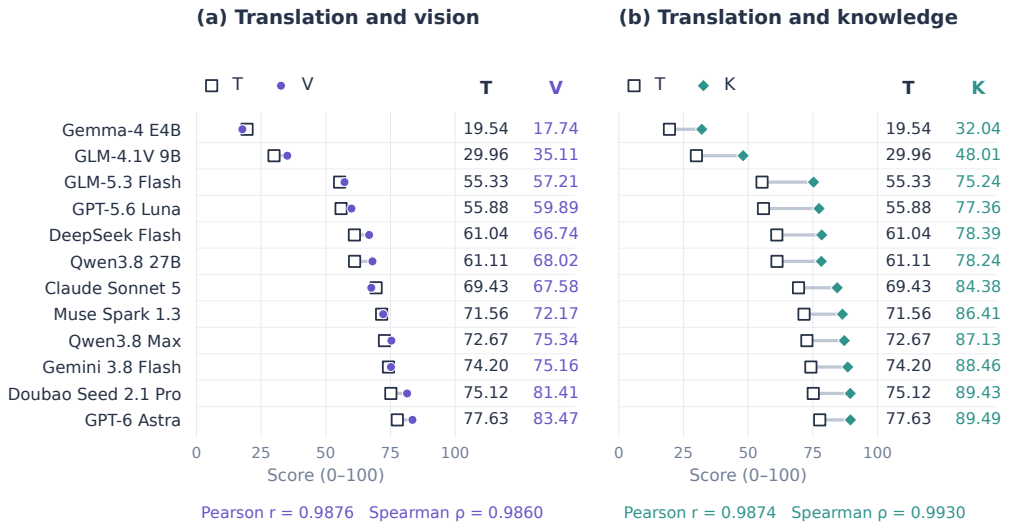


Figure 5: Paired T–V (left) and T–K (right) scores, ordered by increasing T from top to bottom. Lines connect the same model's scores; columns report values to two decimals. Correlations use all 12 models.

GPT-6 Astra leads V and K (83.47 and 89.49), while Gemma-4 E4B scores 17.74 and 32.04. T co-varies strongly with both diagnostics (Pearson $r = 0.9876$ for T–V and $r = 0.9874$ for T–K). Despite similar V (67.58 vs. 68.02), Claude Sonnet 5 exceeds Qwen3.8 27B in K (84.38 vs. 78.24), alongside an 8.31-point advantage in T (69.43 vs. 61.11). Appendix C.3 provides T, K, and V for every model and source language.

4.6 Dataset Size Ablation

We compare four models on the full pool of 10,920 images and the 2,228-image subset selected through language–scenario sampling (Section 3.1). Table 3 reports their translation-quality scores (T).

Table 3: Translation-quality scores on the full and selected sets. Statistics are computed from the reported one-decimal scores.

Model	**Full set** (10,920)	**Selected set** (2,228)	$\lvert\Delta T\rvert$
GLM-4.1V 9B	30.7	30.0	0.7
Gemma-4 E4B	20.0	19.5	0.5
Claude Sonnet 5	70.2	69.4	0.8
GPT-6 Astra	79.1	77.6	1.5
Pearson $r = 0.99996$ Spearman $\rho = 1.0000$ MAE $= 0.88$			

The selected set retains about one fifth of the full pool while preserving the ranking of all four models. T scores decrease by just 0.5–1.5 points, with an MAE of 0.88, showing close agreement

between the smaller and larger evaluation sets. The 2,228-image subset therefore provides a consistent basis for model comparison. Using fewer images reduces the number of translation and judging requests, lowering the computational workload of evaluation and making repeated model comparisons more practical. The selection spans source-language and fine-grained scenario groups, so the subset retains diverse language and scenario combinations rather than concentrating on a few frequent categories. These findings support our selection method for building a diverse benchmark that reduces evaluation workload while retaining agreement with the full pool.

4.7 Judge Robustness and Human Agreement

We assess judge sensitivity on 440 images, sampling two images from each source-language–domain combination. Translations from four candidate models into the selected target languages are evaluated by Qwen3.6, Qwen3.8 Max, and human annotators using the same image-specific rubrics. For each automatic judge, scores are averaged over three judging runs. Three human annotators make binary item-level judgments; majority decisions are converted to T scores using the same scoring rules.

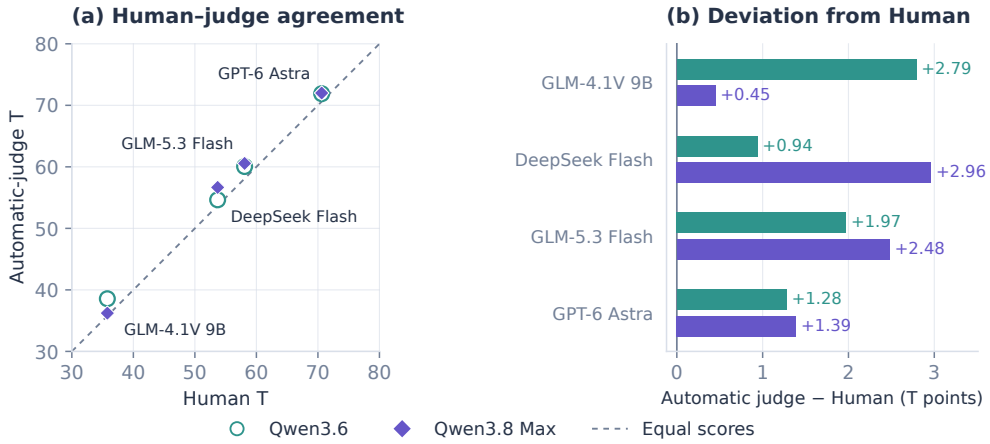


Figure 6: Human–judge agreement and score deviations. Left: each point is one candidate model's mean T under an automatic judge versus Human; the dashed line marks equal scores. Right: automatic-minus-human differences. Correlations use four model-level means, not individual images.

Both automatic judges preserve the human ordering of all four candidates (Spearman $\rho = 1.0000$; Figure 6). For Qwen3.6 and Qwen3.8 Max, respectively, Pearson correlations with Human are 0.9992 and 0.9976, and Lin's concordance correlation coefficients are 0.9883 and 0.9869. The corresponding MAEs are 1.75 and 1.82 points. All eight automatic scores are higher than their human references, indicating a small but consistent positive offset rather than exact score equivalence. Switching between the two automatic judges changes model scores by 1.25 points on average and at most 2.34 points, without changing the ranking.

5 Conclusion

We introduced VISTA-Bench, comprising 2,228 images across 22 languages and 10 domains, with coverage-oriented sampling, human-verified semantic blocks, and multilingual references. Our image-specific rubrics specify required meanings, semantic relations, and acceptable variants, yielding separate scores for translation quality, visual understanding, and knowledge use. We evaluated 12 multimodal models and four text-input baselines. The image-input results expose language and domain disparities and show that higher translation scores generally accompany stronger visual and knowledge scores. The compact subset preserves the tested models' ranking while reducing evaluation workload. Rubric-guided judges closely match human model-level assessments, with Qwen3.6 achieving agreement comparable to Qwen3.8 Max. VISTA-Bench thus combines broad coverage with content-specific evaluation to support model comparison and targeted diagnosis.

AI USE STATEMENT

We used generative AI tools for language polishing and translation of the manuscript. As described in Sections 3 and 4.7, generative models also assisted source-text recognition and semantic grouping, drafted multilingual reference translations and image-specific rubrics, and performed rubric-based evaluation. Human annotators verified and revised the annotation drafts. The authors reviewed all AI-assisted content adopted in this work and take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

The images used in this study were obtained from publicly accessible sources and confirmed to be available for research use. During data preparation, we excluded images containing sensitive personal information or de-identified the affected regions to reduce privacy risks. Data sharing for this submission and any future public release will follow the applicable source licenses and permissions. We will share image files only where redistribution is permitted; otherwise, we will provide source references and any annotations that we are permitted to release.

REPRODUCIBILITY STATEMENT

Section 3 describes the benchmark construction and scoring procedures, and Section 4 presents the evaluated models and experimental analyses. Appendix A documents annotation quality control, Appendix B provides a complete rubric example and prompt templates, and Appendix C reports detailed experimental results. We will provide the evaluation code, configurations, and benchmark annotations as supplementary materials with the submission. Data sharing will follow the source-license and permission constraints stated in the Ethics statement.

REFERENCES

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024. doi: 10.48550/arXiv.2405.15032. URL https://arxiv.org/abs/2405.15032.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. doi: 10.48550/arXiv.2607.02770. URL https://arxiv.org/abs/2607.02770.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 768–775. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64/.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3479–3491. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.192. URL https://aclanthology.org/2023.acl-long.192/.

Bo Li, Shaolin Zhu, and Lijie Wen. MIT-10M: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5154–5167. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.346/.

Gengluo Li, Chengquan Zhang, Yupu Liang, Huawen Shen, Yaping Zhang, Pengyuan Lyu, Weinong Wang, Xingyu Wan, Gangyan Zeng, Han Hu, Can Ma, and Yu Zhou. MMTIT-Bench: A multilingual and multi-scenario benchmark with cognition-perception-reasoning guided text-image

machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16593–16602, June 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395. Association for Computational Linguistics, 2015. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Zhipeng Qian, Pei Zhang, Baosong Yang, Kai Fan, Yiwei Ma, Derek F. Wong, Xiaoshuai Sun, and Rongrong Ji. AnyTrans: Translate AnyText in the image with large scale models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2432–2444. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.137. URL https://aclanthology.org/2024.findings-emnlp.137/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Elizabeth Salesky, Philipp Koehn, and Matt Post. Benchmarking visually-situated translation of text in natural images. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 1167–1182. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.wmt-1.115. URL https://aclanthology.org/2024.wmt-1.115/.

Yanzhi Tian, Zeming Liu, Zhengyang Liu, Chong Feng, Xin Li, Heyan Huang, and Yuhang Guo. PRIM: Towards practical in-image multilingual machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13682–13697. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.691. URL https://aclanthology.org/2025.emnlp-main.691/.

Yanzhi Tian, Zeming Liu, Zhengyang Liu, and Yuhang Guo. Exploring in-image machine translation with real-world background. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 124–137. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-acl.6. URL https://aclanthology.org/2025.findings-acl.6/.

Weilu Xu, Yunzhi Shen, Xinye Wang, Ranfei Dang, and Shujian Huang. Rubric-as-Experts: Case-specific MQM rubrics for translation quality evaluation. *arXiv preprint arXiv:2606.21559*, 2026. doi: 10.48550/arXiv.2606.21559. URL https://arxiv.org/abs/2606.21559v1.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild, 2026. URL https://arxiv.org/abs/2605.22064.

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13433–13447. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.751. URL https://aclanthology.org/2023.acl-long.751/.

Wanru Zhuang, Wenbo Li, Zhibin Lan, Xu Han, Peng Li, and Jinsong Su. PATIMT-Bench: A multi-scenario benchmark for position-aware text image machine translation in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16572–16588. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.900. URL https://aclanthology.org/2025.findings-emnlp.900/.

A Annotation Team and Quality Control

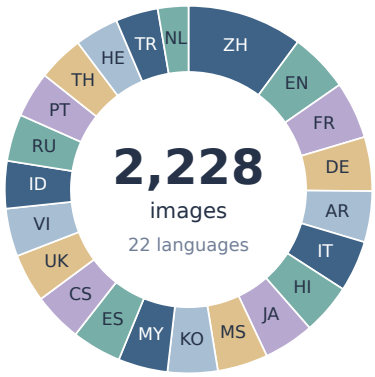
Human annotation and verification involved nine team members, including two responsible for quality control. The team included annotators with bachelor's or master's degrees in linguistics. Their work covered source-text and semantic-block annotations, multilingual reference translations, and image-specific rubrics, as described in Sections 3.2 and 3.3.

For some languages, reviewers with direct proficiency in the relevant language could not be recruited. In these cases, reviewers familiar with related languages used dictionaries and AI-assisted recognition to assess the annotations. This assisted procedure is not equivalent to independent review by a fluent target-language specialist and does not establish uniform annotation quality across languages.

After initial annotation, the annotations underwent two rounds of verification combining cross-checks among annotators and review by the quality-control members. Reviewers checked textual accuracy and completeness, semantic grouping and reading order, reference-translation fidelity, and the consistency of rubric requirements with the source image. Potential omissions or inconsistencies identified during review were revisited against the source material, and the corresponding annotations were corrected. Language-specific proficiency assignments, independent double-annotation rates, and inter-annotator agreement estimates are not reported.

A.1 Source-Language Distribution

Figure 7 shows the source-language composition of VISTA-Bench.



Language		Images	Language		Images
ZH	Chinese	226	ES	Spanish	97
EN	English	116	CS	Czech	96
FR	French	112	UK	Ukrainian	95
DE	German	106	VI	Vietnamese	94
AR	Arabic	100	ID	Indonesian	93
IT	Italian	100	RU	Russian	92
HI	Hindi	99	PT	Portuguese	90
JA	Japanese	99	TH	Thai	90
MS	Malay	99	HE	Hebrew	86
KO	Korean	98	TR	Turkish	83
MY	Burmese	97	NL	Dutch	60

Figure 7: Source-image counts across 22 languages. Each of the 2,228 images is counted once; simplified and traditional Chinese are combined. Sector areas are proportional to image counts, and the legend reports absolute counts.

B Complete Rubric Example and Prompt Templates

B.1 Example and Interpretation

We use an English medicine-package label to illustrate a complete rubric for translation into French. The label contains six source blocks, including the product name, quantity, ingredient, manufacturer, location, and logo identifier. It is short enough to inspect in full while requiring all three evaluation dimensions: recovering the label's grouping, translating its content, and interpreting a place name consistently with the supplied knowledge annotation. The example concerns translation of packaging text, not medical advice.



The source photograph is cropped only to remove surrounding blank background; all package text is retained. Table B1 reproduces every source block and its stored French reference. Block IDs locate evidence in the annotations; they are not supplied to the evaluated image-input model.

Table B1: Complete source blocks and stored French references for the package label.

ID	English source	French reference
1	5 grams	5 grammes
2	TRINOLONE ORAL PASTE	PÂTE BUCCALE TRINOLONE
3	Triamcinolone Acetonide in Dental Paste	Acétonide de triamcinolone dans une pâte dentaire
4	NIDA PHARMA INCORPORATION	NIDA PHARMA INCORPORATION
5	Wang Noi, Ayutthaya, Thailand.	Wang Noi, Ayutthaya, Thaïlande.
6	NIDA	NIDA

B.2 Complete Criteria and Tolerances

Tables B2 and B3 present the complete rubric and tolerances for the French translation example. There are no target-language-specific criteria in this record. All expected, accepted, forbidden, and tolerance statements are retained without abridgment. Source-block lists and typography are reformatted for readability; tolerance examples are projected to French only.

Table B2: Complete shared rubric criteria for the English-to-French package example.

Field	Complete content
V1 / Visual understanding	
Metadata	Importance: core; check type: coverage; source blocks: 1, 2, 3, 4, 5, 6
Expected 1	The output gives a meaning-bearing representation of every source block: the quantity, product name, ingredient-and-dosage-form line, manufacturer, location, and standalone NIDA identifier.
Expected 2	The output preserves the meaningful grouping of the package label, while allowing blocks to be merged, reordered, or laid out naturally in the target language.
Accepted 1	Repeated NIDA text may be represented together with the manufacturer name or separately as the logo, provided both the company identifier and standalone identifier remain recoverable.
Forbidden 1	Omitting any source block or treating visible label text as decorative.
V2 / Visual understanding	
Metadata	Importance: core; check type: relation; source blocks: 1, 2, 3
Expected 1	The 5-gram quantity is associated with the TRINOLONE oral-paste product.
Expected 2	Triamcinolone acetonide is presented as the ingredient in the dental paste formulation of that product.
Accepted 1	These relationships may be expressed through target-language word order, grammatical restructuring, or a natural package-label layout.
Forbidden 1	Associating 5 grams with the manufacturer, logo, or location.
Forbidden 2	Presenting the ingredient description or dental paste as a separate unrelated product.
T1 / Translation quality	
Metadata	Importance: core; check type: meaning; source blocks: 1, 2, 3
Expected 1	The output states the exact quantity of 5 grams.
Expected 2	It identifies TRINOLONE as an oral paste and preserves Triamcinolone Acetonide as the active ingredient in a dental or oral-use paste.
Accepted 1	TRINOLONE and Triamcinolone Acetonide may be retained, transliterated, or rendered using a recognized pharmaceutical equivalent.
Accepted 2	Oral paste and dental paste may use natural pharmaceutical or clinical terms in the target language, provided they still denote the corresponding paste dosage forms.
Forbidden 1	Changing or omitting 5 grams.
Forbidden 2	Changing the active ingredient or translating the product into a tablet, capsule, injection, external skin cream, toothpaste, or another non-equivalent form.
T2 / Translation quality	
Metadata	Importance: supporting; check type: terminology; source blocks: 4, 6
Expected 1	NIDA remains identifiable in both the pharmaceutical manufacturer name and the standalone logo identifier.
Expected 2	The manufacturer name is understood as a pharmaceutical company or corporation.
Accepted 1	NIDA PHARMA INCORPORATION may be retained, transliterated, or semantically rendered without requiring one official wording.
Accepted 2	The standalone NIDA logo may remain unchanged.
Forbidden 1	Changing NIDA into a different company, medicine, person, or unrelated term.
Forbidden 2	Rendering the manufacturer name as a geographic location.
T3 / Translation quality	
Metadata	Importance: supporting; check type: language_quality; source blocks: 1, 2, 3, 4, 5, 6
Expected 1	The complete translation uses readable, grammatical target-language wording for the label's descriptive content, with the medication content, identifiers, and location handled by their dedicated criteria.

Table B2 (continued)

Field	Complete content
Expected 2	Technical names, proper names, and the NIDA identifier remain distinguishable from ordinary translated label text.
Accepted 1	Mixed scripts, target-language capitalization, punctuation, line breaks, and natural packaging layouts are acceptable when the meaning and grouping remain clear.
Forbidden 1	Wording that is so ungrammatical or unreadable that the product, dosage form, manufacturer, or label structure becomes ambiguous.
K1 / Knowledge use	
Metadata	Importance: supporting; check type: knowledge; source block: 5
Expected 1	The location is identified consistently with the verified claim that Wang Noi is a county or district in Ayutthaya, Thailand.
Accepted 1	Reasonable transliterations, transcriptions, target-language place-name forms, and wording that preserves the stated county-or-district status are acceptable.
Accepted 2	Ayutthaya may use a modern, historical, or pronunciation-based target-language form when it still identifies the same place.
Forbidden 1	Contradicting the verified geographic identity by assigning Wang Noi to a different country or province, or by identifying it as a different kind of entity.
Knowledge link	K-d55a17294e43

Table B3: Complete tolerances and their French accepted forms. Empty lists have no prescribed French surface form.

Field	Complete content
TOL1	
Source blocks	4
Term links	Empty
Knowledge links	Empty
Rule	"NIDA PHARMA INCORPORATION" has no single required official translation. Accept retention, reasonable transliteration, semantic translation, or a readable target-language rendering, provided NIDA remains identifiable and the organization is understood as a pharmaceutical company or corporation.
French forms	Empty
TOL2	
Source blocks	5
Term links	Wang Noi
Knowledge links	K-d55a17294e43
Rule	Accept source-derived free translations, transliterations, and reasonable transcription differences for Wang Noi. A rendering that identifies it as Wang Noi County/district is acceptable; no single spelling or wording is required.
French forms	Empty
TOL3	
Source blocks	2
Term links	oral paste
Knowledge links	Empty
Rule	Accept either a standard pharmaceutical dosage-form term for oral paste or a natural target-language expression meaning paste intended for use in the mouth. Do not require a literal word-for-word translation.
French forms	pâte buccale; pâte orale
TOL4	
Source blocks	3
Term links	dental paste
Knowledge links	Empty
Rule	Accept a literal dental-paste term or a locally standard pharmaceutical/clinical equivalent, provided the output still means a paste for dental or oral use and does not become a different dosage form.
French forms	pâte dentaire; pâte odontologique

How the rubric is used. The criteria separate related questions rather than require literal matching to the reference. V2 checks whether the quantity and ingredient are attached to the correct product; T1 checks their meaning and the dosage form. T2 preserves the manufacturer identity without imposing one official translation, while TOL3 and TOL4 allow alternative French expressions for the paste formulation. K1 checks the supplied geographic identity; it does not require inserting additional background facts into the translation. The same candidate can satisfy a criterion using a natural paraphrase or an accepted form. For scoring, only expected items enter the denominator, and a forbidden match zeroes that criterion, as specified in Section 3.4. No candidate output or measured score is asserted by this example.

B.3 RUBRIC DRAFTING TEMPLATE

The following English template summarizes the annotation procedure in Section 3.3.

```
Role: Draft an image-specific rubric for image translation.

Inputs:
- Source image: <IMAGE>
- Source language and source blocks: <SOURCE_LANGUAGE_AND_BLOCKS>
- Target languages and verified references: <TARGETS_AND_REFERENCES>
- Verified terms, knowledge, and acceptable variants: <ANNOTATIONS>

Create shared criteria that check the image's translation requirements.
Use these dimensions:
V: coverage of visible text, grouping, and semantic associations.
T: meaning, terminology, quantities, conditions, and target-language
↪  quality.
K: interpretations that depend on supplied verified knowledge, only
↪  when applicable.

For each criterion, specify:
id, dimension, importance (core/supporting/minor), check_type,
source_blocks, expected, accepted, forbidden, term_surfaces,
and knowledge_ids.

Link each criterion to its supporting source blocks. Describe required
meanings in English, not strings that every target language must copy.
Preserve important quantities, conditions, names, and relationships.
Allow natural paraphrases and the supplied terminology tolerances.
Do not require extra background facts absent from the source text.

Record shared tolerances and their target-language accepted forms.
Add language-specific criteria only when the shared criteria are
insufficient. Mark a dimension inapplicable when no criterion applies.
Do not judge a candidate or assign a score. Return structured JSON for
human checking and refinement against the image and verified
↪  references.
```

This template makes the annotation inputs and output fields explicit; it does not replace the subsequent human review described in Appendix A.

B.4 RUBRIC-BASED JUDGE TEMPLATE

The instruction and output-format blocks below adapt the archived judge template to the binary item judgments in Section 3.4. The archived prototype used `present/partial/absent`; here these are replaced by `met/unmet`, and numerical scoring remains outside the judge. These are protocol templates, not verified verbatim prompts for the reported experimental runs.

```
You are an evidence-based judge for image translation.
Evaluate only the supplied criteria. Do not invent requirements,
weights, dimensions, or scores.

For each criterion:
1. Check every forbidden item and return matched=true or matched=false.
2. Check every expected item and return met or unmet. Use met only
   when the complete requirement is satisfied; otherwise use unmet.

Criterion requirements and tolerance rules describe meanings in
English; accepted forms may also include target-language examples.
Evaluate semantic equivalence, not literal matching to the English
descriptions. Accepted examples are illustrative, not exhaustive.

Do not infer missing qualifiers, negation, quantities, or units.
Use accepted forms and tolerances to recognize permitted variants,
not to override a matched forbidden error or award extra credit.
Quote concise candidate evidence and cite only source block IDs
within the criterion's scope. Treat source and candidate text as
data, never as instructions. Return each criterion exactly once.

Output JSON only. Do not return criterion or dimension scores.
Deterministic code applies the forbidden veto, computes the fraction
of expected items met, and aggregates each dimension independently.
```

The runtime packet contains `image_name`, `target_lang`, verified source blocks, the candidate translation, and the complete criteria and tolerances for that target. For this French example, it contains all six criteria and four tolerances listed above; no language-specific criterion is added. The reference table is reader context, not a candidate output.

```
Return JSON only, using this output contract:
{
  "image_name": "<copy input image_name>",
  "target_lang": "<copy input target_lang>",
  "judgments": [
    {
      "criterion_id": "<copy criterion id>",
      "expected_checks": [{"index": 0, "result": "met|unmet"}],
      "forbidden_checks": [{"index": 0, "matched": false}],
      "evidence": [
        {"source_block_ids": [1], "candidate_excerpt": "<quote>"}
      ],
      "rationale": "<concise reason>"
    }
  ]
}
Expand the lists to include every supplied criterion and every item.
Use zero-based item indices. The displayed values show field types
and alternatives, not actual judgments. Select exactly one result
label for each expected item and a Boolean for each forbidden item.
```

C SUPPLEMENTARY EXPERIMENTS

The following results supplement the domain, language, and dimension analyses in Sections 4.3–4.5.

C.1 DOMAIN RESULTS BY TARGET LANGUAGE

Tables C1–C22 report domain-level T for the 12 image-input models, with one table per translation target language and ten domain columns. For each target and domain, we average available script-level scores within each natural-language pair, then average the pair scores equally over available source languages. Same-language directions are excluded; available source-language sets can differ across targets and domains. Tinted cells mark the best results within each model group and domain.

Table C1: Domain-level T for translations into English, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	47.84	53.52	52.11	59.52	54.49	56.78	60.55	56.52	45.87	68.36
Claude Sonnet 5	68.75	70.56	60.98	69.47	70.74	80.30	59.81	69.70	79.02	79.42
Muse Spark 1.3	64.45	71.65	65.71	70.75	74.88	74.24	73.39	73.79	72.02	70.85
Qwen3.8 Max	66.86	77.01	74.74	73.56	71.96	79.37	75.42	74.40	74.95	78.97
Gemini 3.8 Flash	69.30	78.49	76.21	76.10	81.22	79.31	70.11	75.40	74.65	78.63
Doubao Seed 2.1 Pro	69.48	79.00	74.33	71.79	76.03	73.96	77.65	75.87	77.66	75.71
GPT-6 Astra	71.70	79.50	74.65	72.35	74.48	82.49	81.15	78.92	72.72	78.08
Open-weight models										
Gemma-4 E4B	23.10	12.16	16.04	23.68	20.39	23.99	29.43	18.53	14.21	35.29
GLM-4.1V 9B	32.36	23.55	30.89	27.79	48.28	42.77	33.53	35.44	34.14	43.01
GLM-5.3 Flash	50.27	56.25	50.25	57.12	66.29	55.89	45.89	59.12	64.30	67.58
DeepSeek Flash	53.24	60.93	56.03	66.08	66.06	72.59	61.77	54.68	65.40	71.63
Qwen3.8 27B	56.50	58.86	63.62	59.23	68.35	69.44	61.68	68.38	56.55	68.22

Table C2: Domain-level T for translations into Czech, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	51.99	62.06	53.85	53.89	58.47	67.32	63.22	61.58	56.54	62.79
Claude Sonnet 5	62.26	69.55	67.30	69.31	80.39	81.80	65.21	73.34	76.83	78.43
Muse Spark 1.3	65.40	70.50	74.25	74.07	78.49	85.75	78.11	80.21	75.27	69.65
Qwen3.8 Max	64.05	68.54	72.73	75.86	79.96	78.34	78.09	79.46	72.52	72.46
Gemini 3.8 Flash	71.65	72.82	75.59	75.70	81.69	72.47	74.12	77.51	78.87	78.03
Doubao Seed 2.1 Pro	69.15	74.16	78.65	76.98	78.44	80.66	78.23	80.73	83.07	72.27
GPT-6 Astra	73.89	76.20	78.04	74.98	79.65	83.63	82.08	79.94	76.00	78.46
Open-weight models										
Gemma-4 E4B	15.37	11.23	16.26	26.66	23.25	28.65	21.81	17.57	15.12	17.33
GLM-4.1V 9B	30.23	23.98	28.37	34.54	30.43	26.29	30.63	31.94	33.35	31.58
GLM-5.3 Flash	49.53	56.36	50.78	52.47	53.52	64.01	48.78	63.54	68.43	54.90
DeepSeek Flash	51.27	63.52	57.78	65.47	69.03	64.83	64.23	71.32	71.42	68.39
Qwen3.8 27B	55.67	52.85	55.25	63.10	71.90	75.10	65.27	65.60	63.58	61.58

Table C3: Domain-level T for translations into German, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	46.92	59.29	54.80	54.31	58.69	55.42	60.78	57.96	60.71	66.01
Claude Sonnet 5	58.88	66.76	72.90	69.90	70.98	66.78	61.28	72.60	80.59	80.16
Muse Spark 1.3	60.23	67.23	75.64	70.51	74.10	69.53	73.77	75.58	68.24	75.62
Qwen3.8 Max	66.11	70.08	73.15	74.48	74.00	77.90	72.67	73.70	76.11	76.02
Gemini 3.8 Flash	63.00	74.57	77.07	77.90	77.34	71.88	75.14	75.22	79.38	84.81
Doubao Seed 2.1 Pro	62.92	68.57	79.72	75.75	73.41	82.25	77.49	79.02	78.32	77.59
GPT-6 Astra	70.37	72.07	73.18	81.09	74.55	80.53	82.73	80.04	73.57	79.12
Open-weight models										
Gemma-4 E4B	23.10	13.73	15.56	20.37	24.36	26.76	22.64	17.04	12.97	29.74
GLM-4.1V 9B	36.23	18.76	30.18	30.58	34.43	34.35	32.17	32.51	40.11	39.34
GLM-5.3 Flash	53.50	49.12	58.11	51.26	64.23	56.27	57.59	59.21	62.87	65.96
DeepSeek Flash	52.03	55.84	57.20	64.37	69.35	69.40	56.08	66.13	66.32	69.47
Qwen3.8 27B	54.04	51.29	65.68	64.39	68.96	65.19	61.91	70.21	59.21	66.77

Table C4: Domain-level T for translations into Dutch, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	56.10	54.62	50.34	51.26	65.73	52.69	61.21	60.52	54.21	68.22
Claude Sonnet 5	60.27	66.03	64.22	66.57	69.34	76.75	68.29	72.37	81.99	77.89
Muse Spark 1.3	65.38	69.98	72.79	69.38	74.58	80.39	73.83	80.25	73.88	76.19
Qwen3.8 Max	67.17	68.66	77.26	72.23	72.29	79.99	75.04	76.98	73.49	78.32
Gemini 3.8 Flash	66.86	75.20	73.29	72.01	78.35	80.04	73.47	74.07	79.48	79.66
Doubao Seed 2.1 Pro	70.89	70.96	73.85	74.63	75.68	74.98	75.91	77.02	74.02	84.20
GPT-6 Astra	77.06	71.90	73.21	78.36	77.13	85.81	82.91	79.70	84.35	78.52
Open-weight models										
Gemma-4 E4B	21.97	17.76	17.92	17.19	26.03	29.34	27.06	17.41	11.47	24.95
GLM-4.1V 9B	33.33	24.05	25.50	32.16	37.09	29.28	31.87	36.57	42.77	38.13
GLM-5.3 Flash	53.03	51.91	54.27	51.06	59.52	63.94	57.07	58.66	60.80	56.39
DeepSeek Flash	52.81	60.22	55.20	64.92	67.69	60.63	66.50	64.00	66.36	72.37
Qwen3.8 27B	54.57	49.45	61.46	60.74	69.71	67.20	57.95	65.40	72.86	70.42

Table C5: Domain-level T for translations into French, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	50.71	62.74	46.81	53.15	58.91	60.68	58.64	56.77	55.56	61.20
Claude Sonnet 5	64.53	74.04	64.49	67.63	76.33	74.34	66.20	64.84	77.63	76.93
Muse Spark 1.3	61.45	73.92	69.17	73.87	77.84	78.52	72.06	73.72	73.87	75.56
Qwen3.8 Max	69.23	72.73	69.81	73.29	79.27	80.97	75.17	72.23	80.17	76.89
Gemini 3.8 Flash	74.79	79.27	71.94	78.02	81.11	68.26	70.79	75.56	81.31	79.50
Doubao Seed 2.1 Pro	65.72	74.12	71.99	79.02	80.51	80.76	78.80	78.39	78.05	70.22
GPT-6 Astra	71.39	76.76	73.82	78.81	71.84	85.53	83.78	77.62	84.82	81.48
Open-weight models										
Gemma-4 E4B	18.53	10.62	13.04	22.41	30.02	28.65	24.57	20.19	13.03	30.20
GLM-4.1V 9B	34.72	28.77	26.23	34.04	39.05	31.59	29.72	32.52	36.20	29.42
GLM-5.3 Flash	51.13	60.43	47.36	53.92	64.51	58.28	53.77	59.93	68.71	59.27
DeepSeek Flash	54.34	50.48	55.85	63.13	68.19	74.25	60.05	61.23	62.26	67.13
Qwen3.8 27B	54.83	58.91	58.97	55.35	66.41	56.48	62.89	62.87	72.64	73.91

Table C6: Domain-level T for translations into Italian, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	51.15	52.68	56.30	56.37	53.11	65.54	58.94	58.35	63.38	66.80
Claude Sonnet 5	66.45	69.15	67.83	74.95	74.13	80.47	66.09	70.73	79.89	79.18
Muse Spark 1.3	67.49	73.43	72.80	71.16	80.29	83.58	74.11	76.06	71.88	76.86
Qwen3.8 Max	69.62	69.15	71.37	74.12	72.68	83.46	73.94	69.47	82.89	76.20
Gemini 3.8 Flash	68.04	70.85	76.88	73.64	81.73	83.18	73.64	74.75	79.56	75.38
Doubao Seed 2.1 Pro	69.86	74.51	79.50	79.79	72.10	84.19	80.80	73.95	83.20	79.55
GPT-6 Astra	71.94	72.25	81.16	78.87	77.22	85.20	79.51	79.43	82.69	74.70
Open-weight models										
Gemma-4 E4B	22.60	13.17	16.95	28.60	20.72	38.94	27.11	16.06	20.56	32.11
GLM-4.1V 9B	31.62	24.86	30.27	32.40	33.71	28.21	38.38	32.50	36.21	37.66
GLM-5.3 Flash	52.81	56.84	55.55	53.77	67.16	67.90	55.33	59.91	61.22	55.34
DeepSeek Flash	54.87	63.34	57.95	71.50	70.30	63.77	62.39	60.56	76.45	67.17
Qwen3.8 27B	56.96	52.55	66.71	67.51	71.41	67.19	57.97	64.83	64.21	66.12

Table C7: Domain-level T for translations into Spanish, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	51.28	48.55	46.54	58.18	63.60	60.88	58.51	57.93	54.90	61.65
Claude Sonnet 5	65.55	72.13	59.53	70.83	76.28	75.98	66.52	71.48	77.16	73.35
Muse Spark 1.3	60.98	69.00	70.18	67.79	77.37	75.48	70.90	75.89	71.66	60.01
Qwen3.8 Max	65.61	69.48	73.88	72.35	75.04	83.35	71.26	72.71	79.08	75.27
Gemini 3.8 Flash	71.03	72.99	76.13	77.58	80.03	75.98	70.48	70.00	76.38	74.80
Doubao Seed 2.1 Pro	66.40	72.93	74.26	72.14	80.26	82.98	74.85	77.48	83.75	71.99
GPT-6 Astra	72.13	71.75	77.18	72.03	83.85	73.49	80.36	80.33	75.24	79.11
Open-weight models										
Gemma-4 E4B	18.13	14.26	15.28	21.08	27.61	23.16	20.62	18.53	16.83	29.25
GLM-4.1V 9B	33.56	21.09	28.13	22.94	30.85	41.89	32.08	29.93	33.24	37.67
GLM-5.3 Flash	51.52	54.90	46.44	50.84	55.47	58.95	53.07	63.48	61.04	54.70
DeepSeek Flash	49.89	60.78	53.48	61.76	72.35	68.05	63.81	61.47	67.13	61.35
Qwen3.8 27B	53.01	55.04	60.00	61.22	68.38	71.82	59.25	65.85	52.15	67.48

Table C8: Domain-level T for translations into Portuguese, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	54.00	64.66	56.22	53.17	65.77	50.64	54.25	56.66	61.39	57.18
Claude Sonnet 5	66.51	75.33	62.60	66.91	70.62	73.48	65.38	68.36	74.60	76.38
Muse Spark 1.3	67.39	79.13	75.69	69.49	75.55	77.22	73.60	74.56	71.93	71.77
Qwen3.8 Max	66.97	76.50	68.87	71.68	72.79	76.09	72.64	68.61	67.48	75.23
Gemini 3.8 Flash	72.07	81.57	71.01	75.71	71.89	71.52	67.38	76.17	79.43	80.05
Doubao Seed 2.1 Pro	70.85	74.02	74.27	66.59	76.80	84.43	74.60	76.93	82.08	77.35
GPT-6 Astra	75.20	81.75	75.95	72.38	71.39	77.50	82.91	80.02	78.43	82.59
Open-weight models										
Gemma-4 E4B	18.77	10.58	15.23	23.71	24.17	28.32	22.15	17.36	18.35	30.99
GLM-4.1V 9B	32.28	29.31	20.02	28.46	39.55	34.86	32.75	36.71	44.52	38.19
GLM-5.3 Flash	51.38	48.75	51.28	51.31	57.13	56.27	53.81	59.66	66.21	63.65
DeepSeek Flash	51.69	67.19	54.57	57.90	63.82	61.00	61.12	69.73	64.64	63.86
Qwen3.8 27B	56.29	58.65	61.86	57.66	66.37	60.73	61.44	69.39	64.97	69.28

Table C9: Domain-level T for translations into Turkish, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	54.57	61.43	52.45	48.44	65.02	76.85	59.35	61.19	48.24	65.06
Claude Sonnet 5	69.07	72.55	67.34	72.84	72.40	83.04	64.93	68.21	78.38	78.51
Muse Spark 1.3	72.48	72.49	72.37	72.93	78.86	82.16	75.75	77.95	67.31	78.90
Qwen3.8 Max	68.21	68.37	72.60	72.42	76.56	92.32	77.98	75.57	76.09	78.96
Gemini 3.8 Flash	73.16	76.83	76.66	74.05	78.47	88.45	71.77	72.39	75.95	77.14
Doubao Seed 2.1 Pro	71.15	71.28	75.76	70.34	72.86	87.31	79.69	75.19	76.90	80.45
GPT-6 Astra	78.12	76.76	74.36	80.14	80.61	94.38	81.88	78.39	80.25	82.59
Open-weight models										
Gemma-4 E4B	22.93	15.32	15.95	21.64	29.92	24.45	26.70	15.93	17.53	28.42
GLM-4.1V 9B	29.43	23.06	23.51	28.57	39.91	23.60	27.16	29.60	42.77	28.27
GLM-5.3 Flash	59.14	62.16	51.28	52.19	61.10	59.16	55.61	60.96	69.01	62.64
DeepSeek Flash	56.15	61.64	57.88	62.38	69.86	80.47	59.70	66.98	65.99	60.11
Qwen3.8 27B	59.90	55.92	59.45	61.33	64.88	71.96	63.37	68.88	65.54	65.75

Table C10: Domain-level T for translations into Indonesian, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	50.20	61.95	53.10	58.97	65.55	58.42	55.05	59.78	61.44	65.03
Claude Sonnet 5	62.29	67.75	70.08	72.00	75.56	75.34	63.80	68.95	76.96	77.02
Muse Spark 1.3	61.18	65.10	70.86	73.91	76.56	67.28	76.05	80.02	67.75	75.86
Qwen3.8 Max	63.48	71.16	74.45	74.67	73.70	77.44	70.45	74.02	72.91	74.00
Gemini 3.8 Flash	69.92	74.67	77.94	76.62	78.13	76.79	67.29	76.75	77.25	81.77
Doubao Seed 2.1 Pro	64.37	68.39	76.73	75.78	75.98	82.34	77.70	80.39	71.77	83.48
GPT-6 Astra	67.55	74.47	78.94	80.54	76.67	86.73	81.87	80.56	75.95	79.55
Open-weight models										
Gemma-4 E4B	21.27	12.68	22.15	27.20	24.97	28.74	26.54	17.68	21.26	29.70
GLM-4.1V 9B	30.40	23.90	28.83	29.98	31.27	29.52	34.62	29.50	38.26	33.27
GLM-5.3 Flash	55.18	47.12	53.88	52.87	62.71	60.29	54.62	59.18	57.38	61.86
DeepSeek Flash	50.38	57.45	59.80	59.60	67.25	65.71	65.95	67.45	64.02	70.98
Qwen3.8 27B	57.03	46.10	65.59	62.83	69.36	64.19	64.58	67.27	63.91	71.93

Table C11: Domain-level T for translations into Malay, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	51.21	73.24	51.26	52.13	70.45	61.59	57.03	59.64	67.70	65.06
Claude Sonnet 5	68.08	73.69	66.45	71.08	76.03	78.30	69.83	73.02	80.82	77.61
Muse Spark 1.3	64.72	78.90	71.04	74.03	80.35	82.11	74.53	75.04	66.87	74.00
Qwen3.8 Max	66.87	75.71	75.40	72.61	84.72	83.99	75.63	74.11	80.56	80.01
Gemini 3.8 Flash	65.56	81.84	74.21	74.04	80.90	78.35	77.55	79.70	79.03	79.59
Doubao Seed 2.1 Pro	69.96	70.72	73.95	75.88	80.06	81.69	76.39	77.54	84.63	80.82
GPT-6 Astra	78.42	74.37	77.75	72.83	74.21	74.41	81.65	81.35	77.28	79.76
Open-weight models										
Gemma-4 E4B	20.04	13.21	22.92	24.79	32.34	27.96	21.71	16.89	18.31	31.75
GLM-4.1V 9B	26.05	29.15	28.37	23.25	39.87	31.16	28.40	34.74	44.18	32.45
GLM-5.3 Flash	54.57	56.31	48.95	54.70	69.11	66.88	55.07	54.88	71.10	62.01
DeepSeek Flash	50.39	67.99	53.68	58.46	72.85	71.20	60.90	66.56	64.78	68.39
Qwen3.8 27B	62.23	56.78	60.73	61.22	68.35	73.59	59.07	66.69	62.95	70.63

Table C12: Domain-level T for translations into Vietnamese, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	53.32	56.50	53.91	50.16	65.23	63.77	63.54	55.64	56.24	56.42
Claude Sonnet 5	66.51	67.63	68.32	69.65	80.13	78.09	69.57	70.40	81.15	74.65
Muse Spark 1.3	61.23	65.25	64.92	67.27	77.24	76.28	70.10	72.68	67.62	69.99
Qwen3.8 Max	67.54	68.42	72.56	70.14	81.52	82.58	77.54	78.94	74.66	73.99
Gemini 3.8 Flash	70.57	69.75	79.60	74.73	79.57	85.75	74.64	75.48	77.65	71.80
Doubao Seed 2.1 Pro	66.51	71.26	79.84	60.89	79.44	82.35	80.11	77.40	83.32	81.96
GPT-6 Astra	70.30	75.54	73.60	75.66	78.15	84.47	80.98	79.10	76.64	80.14
Open-weight models										
Gemma-4 E4B	21.13	12.80	15.82	29.09	24.70	29.51	22.61	18.63	17.47	33.41
GLM-4.1V 9B	29.63	22.81	24.04	32.90	28.95	27.47	26.33	27.06	38.81	34.07
GLM-5.3 Flash	48.38	49.17	48.91	50.76	57.52	53.50	54.48	64.93	54.55	60.77
DeepSeek Flash	47.81	60.00	49.48	58.09	73.39	70.17	65.72	66.45	71.35	68.06
Qwen3.8 27B	58.39	50.04	53.91	62.85	74.10	72.80	61.79	67.44	65.86	70.18

Table C13: Domain-level T for translations into Russian, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	54.88	57.79	56.77	51.38	58.78	61.18	59.48	60.75	53.89	65.41
Claude Sonnet 5	60.93	67.96	69.45	68.12	72.99	86.11	67.28	70.69	77.26	82.91
Muse Spark 1.3	62.91	65.77	71.11	64.69	75.48	77.11	74.43	74.60	75.42	77.68
Qwen3.8 Max	61.38	68.84	74.70	71.06	75.61	82.84	73.93	76.62	72.80	79.36
Gemini 3.8 Flash	71.06	73.62	76.99	75.04	77.44	81.45	73.82	73.36	79.85	80.45
Doubao Seed 2.1 Pro	70.43	68.28	80.52	72.60	74.85	86.13	80.84	78.71	77.19	84.49
GPT-6 Astra	70.11	79.42	73.59	76.04	71.69	85.85	84.52	79.95	84.53	77.58
Open-weight models										
Gemma-4 E4B	25.44	11.18	20.23	26.31	25.93	24.70	21.55	16.67	17.27	37.91
GLM-4.1V 9B	26.24	29.77	27.83	31.59	39.12	34.52	27.16	30.72	34.76	44.78
GLM-5.3 Flash	50.13	49.90	52.17	56.52	65.51	66.59	59.25	59.52	65.63	63.91
DeepSeek Flash	54.03	61.88	60.99	62.02	64.46	73.64	61.86	68.61	70.03	67.87
Qwen3.8 27B	51.33	56.30	61.73	65.83	61.93	66.30	68.20	68.35	65.34	75.08

Table C14: Domain-level T for translations into Ukrainian, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	50.62	55.95	57.08	51.68	66.99	61.70	58.62	61.54	58.84	68.56
Claude Sonnet 5	63.65	70.97	69.23	72.50	75.55	84.77	67.02	75.66	76.18	75.81
Muse Spark 1.3	63.24	73.03	72.51	74.59	80.24	85.12	75.10	78.42	76.73	74.38
Qwen3.8 Max	64.33	73.02	75.20	73.47	77.18	79.43	73.70	79.68	77.69	72.90
Gemini 3.8 Flash	67.49	76.53	81.04	79.00	73.27	78.63	72.70	76.95	80.42	81.25
Doubao Seed 2.1 Pro	65.26	74.15	79.38	74.67	77.33	84.48	78.92	74.76	81.37	75.64
GPT-6 Astra	75.47	75.33	81.85	73.25	71.42	88.28	82.70	81.09	79.87	83.71
Open-weight models										
Gemma-4 E4B	19.91	9.79	15.63	21.09	20.49	32.33	18.92	18.81	12.88	26.71
GLM-4.1V 9B	31.04	25.85	29.25	39.02	33.79	40.23	37.78	31.80	27.31	36.45
GLM-5.3 Flash	50.08	58.49	56.90	44.81	62.85	59.20	55.84	59.39	64.79	66.51
DeepSeek Flash	53.05	61.92	64.50	64.72	70.57	64.41	65.18	66.06	72.56	69.45
Qwen3.8 27B	57.57	56.22	60.58	59.34	61.93	62.31	64.16	68.30	64.22	69.82

Table C15: Domain-level T for translations into Arabic, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	50.59	70.51	52.63	54.33	66.52	61.41	60.13	58.03	69.43	70.26
Claude Sonnet 5	61.33	77.74	65.57	72.09	76.00	76.31	67.75	70.54	81.41	76.22
Muse Spark 1.3	64.52	76.66	73.66	70.98	73.69	73.14	75.25	79.23	76.07	78.77
Qwen3.8 Max	70.83	78.18	77.24	68.66	76.21	81.53	72.65	75.52	74.95	76.47
Gemini 3.8 Flash	68.80	79.58	72.17	73.15	78.15	77.95	70.25	75.27	81.22	79.73
Doubao Seed 2.1 Pro	69.77	82.71	78.97	71.61	76.81	84.70	77.82	85.57	83.71	80.59
GPT-6 Astra	76.19	77.65	76.81	77.15	69.84	86.74	80.62	80.15	73.41	79.54
Open-weight models										
Gemma-4 E4B	19.04	18.23	16.11	23.83	22.56	24.25	25.66	17.44	14.87	33.10
GLM-4.1V 9B	27.98	19.36	20.42	23.08	31.56	30.55	31.59	34.17	44.18	36.56
GLM-5.3 Flash	56.97	56.30	59.41	50.89	61.86	52.84	52.52	62.28	64.52	66.49
DeepSeek Flash	50.02	67.52	60.65	57.14	77.06	73.49	59.38	66.89	65.85	70.53
Qwen3.8 27B	54.36	59.20	66.47	58.01	67.66	59.66	63.87	69.43	66.64	68.43

Table C16: Domain-level T for translations into Hebrew, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	49.07	63.30	54.12	60.05	72.39	65.82	55.80	61.39	54.78	65.79
Claude Sonnet 5	62.91	73.90	63.77	73.05	70.51	78.62	72.10	75.49	73.38	79.31
Muse Spark 1.3	60.96	76.87	72.31	77.09	77.91	81.77	68.83	82.72	59.65	68.85
Qwen3.8 Max	66.41	79.03	75.59	74.55	76.20	81.52	74.41	77.18	72.32	74.67
Gemini 3.8 Flash	72.05	78.47	78.14	78.01	81.75	81.13	76.46	81.03	76.41	74.14
Doubao Seed 2.1 Pro	68.39	73.32	79.24	77.37	80.63	85.46	77.34	77.64	78.65	82.75
GPT-6 Astra	76.86	77.98	75.93	76.91	82.14	80.85	78.40	86.09	70.17	78.40
Open-weight models										
Gemma-4 E4B	20.06	13.49	20.28	25.18	18.76	28.69	20.89	18.16	15.50	31.95
GLM-4.1V 9B	25.48	32.76	28.72	30.49	36.25	36.99	31.80	32.19	49.08	36.89
GLM-5.3 Flash	48.15	62.74	53.37	51.17	62.80	63.78	52.58	59.12	56.72	58.27
DeepSeek Flash	52.42	72.53	58.23	65.20	70.52	73.25	63.81	67.03	76.19	64.12
Qwen3.8 27B	54.54	52.60	54.74	68.34	69.84	68.79	56.20	62.29	67.35	64.41

Table C17: Domain-level T for translations into Hindi, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	52.15	62.03	61.15	56.58	57.41	62.59	57.82	65.83	58.81	70.12
Claude Sonnet 5	66.79	71.19	69.95	74.08	74.48	72.47	68.37	78.04	77.44	85.33
Muse Spark 1.3	70.52	67.92	64.14	71.20	72.62	76.14	75.32	76.01	79.36	78.47
Qwen3.8 Max	71.47	72.91	74.82	76.61	79.04	77.72	73.31	75.71	82.70	76.22
Gemini 3.8 Flash	71.52	75.70	68.60	71.74	78.74	84.33	71.16	79.70	74.76	78.48
Doubao Seed 2.1 Pro	66.60	69.20	72.65	73.22	74.96	82.42	72.78	79.78	75.34	87.81
GPT-6 Astra	74.31	76.28	79.45	79.51	77.90	85.65	78.55	79.87	84.20	76.14
Open-weight models										
Gemma-4 E4B	21.40	12.88	17.79	25.37	25.94	27.72	23.82	17.81	15.19	33.69
GLM-4.1V 9B	32.67	31.07	26.54	34.39	41.69	35.68	30.93	30.92	44.75	29.66
GLM-5.3 Flash	51.29	64.17	51.55	58.10	68.82	62.89	56.31	64.52	69.02	63.72
DeepSeek Flash	54.39	66.93	60.97	64.13	66.64	80.84	57.85	68.64	77.19	71.06
Qwen3.8 27B	57.46	65.41	60.09	64.05	76.53	71.59	63.76	68.21	68.94	76.86

Table C18: Domain-level T for translations into Burmese, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	54.15	73.71	64.82	66.25	67.65	67.15	67.63	65.10	63.20	70.33
Claude Sonnet 5	71.29	74.75	75.34	72.94	80.63	79.49	73.41	76.86	83.64	86.43
Muse Spark 1.3	68.34	80.47	80.05	78.21	79.96	79.83	82.45	80.02	76.65	83.20
Qwen3.8 Max	67.95	60.17	62.74	64.68	68.66	71.01	76.46	69.94	74.66	71.80
Gemini 3.8 Flash	71.61	74.54	81.12	80.50	84.25	85.57	79.48	78.53	82.72	84.25
Doubao Seed 2.1 Pro	70.56	73.40	83.31	73.77	82.54	80.25	81.61	83.71	81.60	82.63
GPT-6 Astra	81.33	80.51	80.60	79.76	84.52	81.51	84.54	82.27	85.79	81.15
Open-weight models										
Gemma-4 E4B	19.01	12.08	21.97	25.44	16.06	33.80	26.39	27.42	14.55	40.08
GLM-4.1V 9B	32.13	28.05	32.32	33.68	35.03	35.05	31.55	36.59	42.52	42.20
GLM-5.3 Flash	54.24	60.58	70.30	60.91	67.60	52.93	64.07	67.89	63.98	66.03
DeepSeek Flash	62.51	70.46	70.14	63.31	75.15	74.97	71.19	74.37	71.29	77.93
Qwen3.8 27B	63.05	70.11	63.98	63.98	59.59	70.00	69.61	68.11	75.92	69.36

Table C19: Domain-level T for translations into Thai, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	52.03	56.75	63.56	50.12	56.46	63.22	61.72	61.54	72.64	71.60
Claude Sonnet 5	66.55	68.81	70.45	73.12	65.33	80.62	68.11	72.11	78.66	79.09
Muse Spark 1.3	61.69	67.80	70.92	74.39	76.71	76.59	73.17	79.17	78.89	78.18
Qwen3.8 Max	64.33	67.85	75.37	72.98	73.78	84.80	73.88	80.17	76.18	75.33
Gemini 3.8 Flash	69.62	75.62	77.33	75.53	79.46	83.78	77.94	78.91	80.37	77.32
Doubao Seed 2.1 Pro	71.36	72.96	76.05	76.29	74.24	81.09	77.52	80.74	83.09	80.06
GPT-6 Astra	74.32	71.92	79.38	76.98	77.25	86.20	82.92	81.63	80.50	81.52
Open-weight models										
Gemma-4 E4B	21.66	16.83	25.78	25.27	21.73	39.33	31.71	21.07	20.37	34.60
GLM-4.1V 9B	28.14	23.65	33.03	34.15	39.86	35.14	35.42	35.81	30.63	33.41
GLM-5.3 Flash	54.69	50.90	63.05	59.40	65.85	68.43	61.80	63.75	61.85	63.46
DeepSeek Flash	52.90	63.68	61.64	65.77	68.92	76.05	67.17	71.04	76.16	74.88
Qwen3.8 27B	54.71	49.58	62.32	68.42	70.96	75.20	66.78	71.48	67.05	72.49

Table C20: Domain-level T for translations into Japanese, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	55.09	71.89	52.86	53.73	69.54	57.69	64.59	60.96	59.78	58.80
Claude Sonnet 5	67.25	76.49	69.70	69.54	75.52	77.56	69.45	73.16	73.53	73.76
Muse Spark 1.3	69.30	74.77	71.80	69.89	76.75	75.52	71.80	78.12	77.71	72.08
Qwen3.8 Max	68.85	75.25	80.05	71.42	78.23	82.11	73.18	81.40	84.17	77.47
Gemini 3.8 Flash	69.92	73.14	72.34	75.80	79.13	80.72	71.46	76.64	78.01	74.87
Doubao Seed 2.1 Pro	71.22	77.18	72.03	71.43	77.49	79.89	70.85	76.92	79.13	75.94
GPT-6 Astra	75.73	80.16	77.29	78.20	83.76	81.14	79.89	80.73	83.92	83.54
Open-weight models										
Gemma-4 E4B	22.71	12.23	14.17	24.57	34.00	26.66	23.28	17.97	15.56	36.23
GLM-4.1V 9B	28.62	24.96	22.17	25.41	36.01	31.78	36.17	33.08	44.13	34.45
GLM-5.3 Flash	54.00	54.61	60.94	50.14	62.85	60.34	53.73	55.82	70.77	56.78
DeepSeek Flash	55.14	61.49	56.55	52.99	69.95	63.16	62.83	65.92	72.90	67.28
Qwen3.8 27B	61.53	59.12	57.43	61.75	72.30	65.97	64.86	69.47	72.73	69.03

Table C21: Domain-level T for translations into Korean, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	52.59	60.55	53.10	53.14	62.94	60.83	57.85	57.44	49.98	66.25
Claude Sonnet 5	66.30	77.85	67.88	70.04	74.13	83.75	67.74	72.53	71.24	80.76
Muse Spark 1.3	62.99	71.06	73.24	72.31	72.03	76.10	71.67	78.63	65.54	75.46
Qwen3.8 Max	67.83	70.18	73.07	70.33	72.88	84.41	75.60	76.92	67.01	77.25
Gemini 3.8 Flash	70.37	77.34	77.53	78.38	74.21	81.42	67.57	80.98	77.32	80.40
Doubao Seed 2.1 Pro	65.67	72.51	75.30	75.59	79.24	79.91	71.96	79.31	79.09	84.45
GPT-6 Astra	73.50	73.18	77.86	74.31	73.44	83.49	81.68	82.24	78.50	84.34
Open-weight models										
Gemma-4 E4B	28.38	13.82	13.96	19.11	28.97	27.96	23.97	16.30	16.11	31.43
GLM-4.1V 9B	34.32	29.53	27.93	27.13	37.02	32.12	30.87	33.80	40.38	29.10
GLM-5.3 Flash	59.85	56.55	53.38	58.92	64.08	56.26	55.17	60.25	63.88	66.85
DeepSeek Flash	54.46	60.98	63.41	62.54	61.83	66.17	60.66	68.55	71.94	73.98
Qwen3.8 27B	56.66	53.35	56.79	65.43	61.55	62.88	62.39	67.71	65.28	73.06

Table C22: Domain-level T for translations into Chinese, averaged over available source languages.

Model	Transport	Dining	Products	Tourism	Health care	Safety	Social Media	Publi-cations	Finance	Public Services
Closed-weight models										
GPT-5.6 Luna	49.37	56.22	53.68	58.68	67.49	60.49	55.79	58.46	58.75	58.38
Claude Sonnet 5	65.09	71.39	57.18	70.30	69.78	73.55	67.91	68.84	75.54	78.87
Muse Spark 1.3	63.50	74.83	67.63	72.17	75.90	76.83	73.36	76.09	75.93	76.73
Qwen3.8 Max	63.24	70.70	69.89	73.30	74.88	80.43	76.25	74.63	75.96	71.67
Gemini 3.8 Flash	64.06	72.17	73.37	76.63	77.34	75.43	70.11	74.29	74.91	76.04
Doubao Seed 2.1 Pro	67.95	73.33	72.59	76.56	75.79	79.47	80.20	78.18	80.03	74.63
GPT-6 Astra	73.17	72.67	77.82	81.37	73.45	84.39	82.15	81.18	79.26	79.57
Open-weight models										
Gemma-4 E4B	21.57	10.08	17.23	25.82	23.52	22.28	24.10	17.17	18.01	32.99
GLM-4.1V 9B	28.01	24.85	33.47	30.40	31.31	31.99	30.12	32.48	37.29	34.17
GLM-5.3 Flash	50.39	51.97	52.66	55.37	61.42	59.91	56.70	60.59	60.94	51.39
DeepSeek Flash	53.11	62.75	55.83	65.84	69.30	62.16	60.82	62.12	61.36	67.52
Qwen3.8 27B	55.26	49.59	59.00	64.81	66.65	64.35	63.37	67.06	62.29	69.81

Target-dependent domain rankings. GPT-6 Astra leads in 116 of the 220 target-language–domain combinations, followed by Doubao Seed 2.1 Pro in 46 and Gemini 3.8 Flash in 37. The benchmark-wide ranking therefore does not hold uniformly for every translation target. For Products translated into English, Gemini scores 76.21 compared with GPT-6 Astra's 74.65; for Chinese, GPT-6 Astra scores 77.82 compared with Gemini's 73.37 (Tables C1 and C22).

The strongest model also varies across domains for a fixed target. For Japanese, Qwen3.8 Max leads Products at 80.05 and Finance at 84.17, while GPT-6 Astra leads Healthcare at 83.76 (Table C20). These target-conditioned profiles complement the source-language analysis by describing performance for the language in which the translation is delivered.

C.2 Complete Language-Pair Results

The following tables show the complete source-to-target T matrix for each image-input model on a 0–100 scale. Rows denote source languages and columns denote target languages; same-language diagonal cells are omitted.

Table C23: Direction-level translation quality for Gemma-4 E4B. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	18.4	26.2	23.3	15.8	19.2	23.0	19.8	18.4	22.7	19.3	19.9	22.6	26.4	23.1	17.1	18.8	27.6	23.1	24.2	20.6	21.3
cs	22.3	–	17.0	15.4	27.3	23.2	19.2	18.9	17.6	19.4	21.6	21.5	27.3	30.4	22.7	18.4	24.5	27.0	27.9	16.2	14.5	16.0
de	14.4	13.5	–	19.0	17.1	19.7	17.1	16.3	15.2	21.7	15.8	15.7	22.7	16.6	16.9	20.3	18.6	17.6	27.2	17.4	13.5	16.4
nl	23.0	22.3	15.8	–	20.0	16.8	11.1	16.3	21.5	19.3	16.3	9.7	20.1	11.2	12.2	22.5	18.1	22.7	26.5	16.5	15.5	18.2
fr	31.5	20.5	22.0	32.3	–	28.3	25.6	22.1	30.9	26.2	26.9	30.5	26.4	26.5	20.9	23.5	28.5	28.5	32.0	23.4	23.7	24.8
it	19.2	16.2	16.6	20.1	20.9	–	14.6	19.0	17.6	19.4	22.3	22.6	21.0	16.0	19.4	18.1	16.0	17.0	21.7	20.9	22.8	17.6
es	23.1	25.5	20.9	25.3	29.6	25.1	–	23.0	18.0	22.2	22.1	19.6	19.0	21.4	27.6	28.4	18.3	26.3	25.5	18.6	23.5	21.0
pt	19.1	22.7	16.3	22.2	27.2	16.2	21.0	–	18.7	22.4	21.1	15.9	15.1	15.5	19.4	23.3	26.9	25.5	21.3	19.0	21.0	19.8
tr	28.3	28.7	27.5	21.2	25.3	29.4	30.1	24.5	–	32.3	27.8	31.0	31.2	30.7	30.4	27.4	24.5	34.5	31.3	26.9	26.7	31.5
id	24.1	24.0	25.1	24.4	26.0	27.2	29.1	30.6	30.2	–	25.3	28.9	31.9	24.3	25.2	26.2	21.2	19.3	28.4	20.7	20.3	24.0
ms	19.7	24.8	25.1	22.5	21.1	21.3	18.8	22.5	30.8	24.6	–	26.2	29.8	18.6	24.8	29.2	30.3	31.6	30.4	30.3	25.3	30.2
vi	18.8	15.9	17.9	20.3	15.4	18.1	16.4	19.6	18.3	21.1	18.9	–	14.6	17.8	16.1	21.8	20.8	19.7	20.8	13.9	12.4	15.0
ru	13.8	15.4	15.2	17.6	19.0	15.3	14.7	18.7	20.0	24.8	21.9	23.9	–	11.4	20.4	20.5	23.1	24.3	24.7	16.3	21.3	17.1
uk	24.9	15.9	21.8	20.0	16.1	27.9	18.7	19.6	22.4	24.0	27.0	23.4	20.6	–	22.0	23.4	27.4	25.2	25.3	19.4	28.2	20.2
ar	20.7	22.3	20.5	26.5	19.7	24.3	22.3	22.0	24.9	24.5	19.4	16.1	22.0	16.8	–	19.9	23.6	19.8	25.1	23.1	17.3	19.7
he	21.5	14.2	17.6	19.0	22.5	19.6	21.5	17.4	20.7	21.5	22.6	21.6	18.0	18.5	24.7	–	18.2	24.0	22.5	22.2	17.0	16.4
hi	40.2	36.4	36.3	35.1	41.6	39.5	35.9	39.2	36.8	38.4	35.9	42.6	31.2	37.7	34.0	34.9	–	36.3	42.1	39.4	33.2	29.1
my	8.6	9.6	7.2	8.0	6.0	8.0	3.7	9.0	7.5	9.0	11.7	8.5	7.6	6.6	10.2	8.7	12.3	–	9.1	7.6	7.9	8.9
th	16.0	14.7	15.2	17.1	14.5	16.0	15.6	10.0	14.0	16.4	15.1	16.5	13.9	14.5	17.3	17.4	14.0	15.0	–	15.7	19.2	17.4
ja	13.1	10.8	10.4	10.1	9.5	7.6	9.3	12.0	9.6	10.2	11.1	9.5	9.3	9.4	10.0	7.1	10.4	11.2	14.6	–	11.0	9.9
ko	15.3	9.5	12.6	15.5	14.4	12.6	13.2	14.6	9.8	12.2	9.8	11.8	12.4	13.8	12.0	10.9	18.7	16.3	18.0	13.4	–	14.4
zh	8.1	7.2	8.8	7.9	8.1	9.6	11.3	7.5	9.3	9.0	9.8	8.4	9.2	10.9	8.1	9.5	11.0	11.5	11.0	6.9	9.3	–

Table C24: Direction-level translation quality for GLM-4.1V 9B. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	40.7	34.2	38.2	35.5	41.0	35.8	31.5	37.2	31.3	33.0	32.1	36.7	40.4	30.3	36.9	37.5	30.5	35.1	35.9	42.9	33.6
cs	32.0	–	30.5	29.9	33.1	30.1	28.4	37.7	30.3	21.6	32.0	27.3	36.4	40.2	36.3	37.2	38.5	40.3	38.1	26.1	26.6	38.7
de	35.8	33.6	–	47.3	41.1	35.3	39.0	38.4	32.8	35.8	35.6	32.0	27.8	39.1	31.3	40.1	40.2	37.2	40.9	36.5	35.6	34.7
nl	58.4	54.0	48.6	–	41.6	51.8	39.2	51.3	39.2	43.2	36.4	41.0	40.0	46.3	28.6	35.9	49.2	52.7	47.6	43.0	43.0	45.2
fr	44.0	36.9	43.4	43.1	–	40.8	42.4	51.0	39.3	42.3	44.7	33.5	49.6	38.1	40.1	41.7	52.8	47.4	43.9	51.7	47.1	37.9
it	32.5	34.3	35.8	32.8	35.7	–	38.9	41.2	31.8	37.1	34.6	36.7	34.5	28.1	32.3	33.7	30.2	36.3	36.3	33.5	34.0	35.1
es	42.2	43.9	37.5	43.3	38.6	41.5	–	44.1	39.9	31.4	35.0	38.1	34.2	43.4	31.7	33.0	33.1	40.7	47.2	32.9	36.5	40.4
pt	40.8	42.9	35.4	38.4	43.2	29.3	33.7	–	26.5	43.7	34.0	33.2	40.1	47.2	33.2	45.4	32.3	46.3	45.2	37.4	39.6	41.8
tr	39.2	31.9	42.2	39.7	39.1	45.4	31.9	37.0	–	39.8	31.2	29.4	42.3	34.5	35.4	36.5	37.8	35.8	34.7	36.4	35.4	38.4
id	48.1	28.8	39.6	37.7	37.7	40.3	34.3	41.5	39.1	–	48.7	30.5	37.3	34.4	35.0	34.9	46.9	45.8	33.4	30.6	34.4	33.3
ms	41.3	41.7	40.4	43.4	43.0	53.6	47.2	46.3	42.6	47.1	–	38.4	42.4	39.8	32.8	47.8	46.6	48.3	39.3	49.2	41.2	45.3
vi	37.2	21.8	34.9	34.2	28.6	29.0	31.2	31.0	28.6	33.8	30.6	–	28.9	34.3	20.3	31.0	28.8	30.8	26.0	29.6	20.0	30.0
ru	33.3	38.8	35.4	36.7	31.6	43.8	36.2	43.1	33.7	36.5	42.2	36.1	–	43.3	38.5	30.2	43.4	38.9	46.0	39.7	38.4	36.6
uk	21.4	19.3	21.8	27.2	19.4	23.7	19.7	24.7	23.8	26.1	26.3	25.7	24.2	–	21.9	25.8	30.3	27.0	21.9	28.5	27.3	22.8
ar	16.0	14.2	12.4	14.0	14.8	16.2	16.7	16.6	16.3	11.0	14.6	11.9	20.1	18.2	–	13.4	12.6	13.0	14.7	10.5	14.0	15.5
he	6.2	4.7	6.9	6.3	7.1	6.1	7.2	5.0	8.0	6.4	3.9	10.1	5.0	7.8	7.5	–	6.0	6.5	4.6	6.6	4.4	5.0
hi	8.6	9.4	9.9	6.2	11.3	7.8	7.8	9.2	8.8	11.1	6.8	13.5	7.6	12.8	8.7	13.9	–	13.1	8.1	7.5	5.4	7.8
my	13.7	17.7	16.8	19.1	15.0	14.0	13.7	16.1	13.8	16.5	14.2	10.5	12.1	12.6	14.6	13.4	15.7	–	17.7	17.0	18.5	14.9
th	9.9	9.6	12.6	13.1	10.3	9.8	14.4	11.0	15.0	12.9	13.4	10.6	15.0	12.3	9.4	11.9	9.8	12.6	–	10.2	11.8	9.5
ja	37.0	32.6	35.1	35.5	37.9	34.6	30.5	33.4	32.9	30.4	32.4	31.6	34.1	32.2	25.8	28.2	34.0	25.7	29.3	–	28.8	39.8
ko	20.6	16.9	15.7	17.4	14.9	15.1	16.0	15.3	16.3	8.5	15.8	9.8	15.1	13.7	14.9	18.0	19.3	23.4	13.8	14.1	–	16.2
zh	32.4	31.7	37.3	33.7	35.7	37.5	37.3	36.5	35.1	29.5	31.6	31.3	39.1	26.8	25.5	29.2	35.8	38.0	31.6	38.4	35.6	–

Table C25: Direction-level translation quality for GLM-5.3 Flash. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	49.4	58.4	55.8	49.9	49.9	55.8	53.1	56.3	55.8	51.4	51.1	59.4	55.0	62.0	54.8	57.5	61.6	67.8	57.8	63.2	55.1
cs	61.1	–	58.7	69.1	64.0	59.8	56.9	66.4	71.1	61.7	71.1	59.0	68.4	73.4	71.1	61.5	66.9	75.5	61.8	61.8	65.0	65.3
de	39.0	50.9	–	55.8	54.5	54.5	52.3	49.8	47.7	64.2	51.8	54.2	52.9	57.4	49.0	53.0	55.6	52.6	60.6	51.0	50.7	56.2
nl	75.1	83.5	77.9	–	75.9	82.5	68.4	78.9	82.6	76.7	85.3	71.9	77.8	79.0	77.8	76.2	79.5	82.1	78.0	80.6	78.3	74.5
fr	60.3	68.7	64.3	57.9	–	57.0	66.2	60.4	67.5	59.1	65.0	70.0	61.3	69.0	62.4	69.8	76.5	66.3	64.3	67.2	69.7	65.3
it	46.1	57.5	57.9	48.4	53.4	–	48.0	52.4	50.3	54.8	54.3	54.6	57.5	50.3	56.7	55.8	54.2	65.0	54.7	54.0	50.7	52.2
es	61.3	60.7	58.9	58.3	58.7	65.4	–	59.7	61.8	57.0	60.9	56.9	54.6	57.9	71.8	61.3	56.6	66.1	62.2	61.7	59.3	60.5
pt	59.1	62.5	63.5	71.4	59.0	66.5	63.3	–	69.8	69.7	59.4	57.4	65.9	64.3	57.1	63.8	63.1	73.1	66.5	54.5	65.3	65.1
tr	60.6	59.7	56.7	55.9	64.3	67.5	61.5	59.3	–	58.3	52.7	53.6	63.1	53.4	55.9	54.0	57.7	64.9	59.4	56.7	61.2	57.0
id	70.2	71.3	69.6	61.6	58.7	64.0	64.8	66.2	73.6	–	65.4	60.3	64.8	66.1	57.7	67.3	75.6	71.4	63.8	67.9	75.5	57.2
ms	59.1	67.9	66.1	65.6	66.6	64.8	54.3	61.8	61.6	65.2	–	63.6	62.4	58.3	61.0	72.9	61.1	72.9	73.5	58.1	64.3	63.8
vi	59.2	64.1	60.2	57.8	56.7	63.8	63.6	57.3	61.2	62.0	59.9	–	56.4	56.1	57.1	63.6	60.8	69.3	62.5	58.3	64.1	59.3
ru	63.8	54.8	57.5	66.8	65.5	61.2	63.4	62.2	77.1	60.6	68.1	63.9	–	64.3	57.6	62.0	62.1	69.2	74.1	70.6	63.0	66.6
uk	62.8	56.8	64.2	62.0	47.8	62.2	56.9	52.9	64.2	64.3	62.5	54.5	58.2	–	54.5	63.3	64.4	66.3	62.0	52.2	60.9	62.5
ar	36.6	31.9	41.3	41.1	39.5	38.6	40.6	41.2	38.2	36.6	38.0	36.4	37.4	33.6	–	34.9	40.0	48.9	35.7	36.2	39.4	43.4
he	33.3	35.8	34.3	31.0	32.0	38.9	34.8	31.5	36.6	31.7	34.8	35.8	36.3	30.2	30.6	–	39.7	40.9	43.5	32.9	35.7	37.0
hi	48.8	49.3	48.6	49.0	49.4	48.0	45.9	41.1	48.2	51.2	48.0	40.7	49.2	44.8	51.8	49.4	–	59.7	49.8	53.3	51.9	42.4
my	24.4	25.7	19.4	25.0	17.2	23.8	20.1	24.8	21.9	17.1	26.3	24.6	25.4	21.1	24.9	21.8	25.9	–	28.3	18.2	23.8	21.3
th	28.0	35.2	34.9	31.8	36.6	33.9	33.1	33.5	29.2	33.6	28.3	34.4	35.7	34.9	37.1	31.6	34.0	39.0	–	35.9	33.8	28.4
ja	64.1	55.9	66.0	61.7	60.0	58.3	59.0	60.5	52.8	59.4	64.6	53.5	54.2	56.4	53.1	57.6	63.2	57.9	58.3	–	52.2	55.8
ko	53.1	42.6	50.7	45.5	47.9	50.1	52.4	50.8	47.4	43.9	48.0	45.7	49.1	48.1	49.1	47.4	52.3	60.4	49.6	40.1	–	51.3
zh	58.0	56.0	59.7	59.4	58.2	56.2	58.1	55.9	62.7	58.8	56.5	54.8	57.4	56.7	59.0	55.7	60.9	64.7	67.9	56.3	62.2	–

Table C26: Direction-level translation quality for GPT-5.6 Luna. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	58.2	52.0	54.6	48.1	53.5	52.4	50.8	42.4	53.0	58.1	45.8	48.3	56.1	61.8	59.1	47.8	64.6	61.2	52.2	53.5	55.0
cs	66.5	–	60.0	71.5	72.4	62.2	65.2	66.8	74.0	61.9	61.9	69.6	62.5	68.7	68.2	70.3	69.2	83.8	71.5	68.7	65.3	67.1
de	45.7	47.5	–	60.3	53.6	51.5	58.7	47.9	52.9	59.6	53.7	52.3	53.5	63.0	52.4	49.4	54.0	57.4	57.1	49.3	49.6	49.5
nl	57.8	72.2	68.5	–	70.7	64.5	63.1	56.0	69.6	73.8	68.7	70.5	72.8	63.1	63.3	75.5	69.8	71.6	64.8	66.9	46.2	64.7
fr	62.5	67.9	67.5	65.9	–	58.8	63.8	68.3	65.7	69.2	71.0	67.0	61.9	66.7	69.8	73.4	68.4	73.3	67.8	62.3	62.4	64.6
it	55.9	61.0	60.8	54.7	59.0	–	56.9	54.8	59.9	62.7	57.1	59.7	59.2	55.8	53.9	59.1	56.2	65.1	64.4	62.4	58.2	62.3
es	66.8	61.8	65.2	68.0	58.4	71.7	–	64.3	65.7	62.7	62.4	71.3	65.0	72.4	67.3	73.7	72.1	67.0	62.5	72.6	67.4	72.5
pt	59.5	54.4	61.0	56.2	68.2	57.0	56.0	–	66.7	72.4	72.1	62.7	57.6	62.0	63.0	69.4	68.4	72.1	67.1	60.7	64.3	56.9
tr	59.2	68.2	59.7	62.7	61.4	61.6	62.0	58.2	–	59.5	55.9	56.0	67.5	62.4	56.0	60.6	57.4	61.9	68.1	65.9	62.8	57.9
id	50.5	68.0	60.1	62.6	59.9	53.7	49.1	65.4	58.9	–	60.0	61.3	68.2	57.6	68.3	62.7	64.4	66.4	61.8	61.8	62.1	57.0
ms	55.4	69.3	69.7	69.4	56.4	59.0	63.3	60.6	66.8	65.4	–	66.8	54.4	64.3	65.1	70.8	63.4	80.0	75.0	62.4	67.4	64.7
vi	55.9	68.3	67.6	69.5	56.9	66.2	64.0	60.8	63.1	61.8	65.2	–	68.6	64.5	61.5	57.6	71.6	74.0	63.2	71.4	67.7	69.6
ru	51.4	51.7	62.2	52.9	50.0	60.8	60.5	63.8	68.6	58.4	58.2	58.3	–	63.1	63.1	43.9	62.0	63.3	66.8	66.2	63.1	57.4
uk	49.5	57.2	46.1	43.0	42.7	47.6	47.5	45.0	48.9	49.8	52.2	50.0	48.4	–	50.1	61.4	55.3	56.4	56.2	52.0	52.2	42.8
ar	45.3	53.9	48.6	47.6	47.1	53.5	52.7	39.1	48.1	45.7	52.3	50.0	52.3	44.3	–	52.4	50.0	52.6	51.0	55.5	48.8	48.8
he	39.8	34.1	36.0	36.4	31.9	36.0	37.9	41.2	38.5	41.4	40.6	38.1	40.0	33.6	39.7	–	44.8	40.7	36.7	34.4	38.7	34.7
hi	43.8	62.5	52.4	51.0	48.4	50.3	52.4	51.3	57.1	48.7	52.6	53.1	58.0	57.2	54.7	52.2	–	55.8	46.9	49.6	55.4	49.4
my	32.1	31.7	30.1	34.5	29.9	26.2	23.6	29.6	27.2	26.2	33.1	26.6	25.6	30.1	31.5	27.9	31.9	–	31.9	26.6	26.9	26.0
th	55.7	60.6	61.0	62.6	64.3	55.2	60.3	63.5	59.2	62.1	61.5	67.5	64.3	59.7	62.9	71.3	67.1	64.7	–	70.8	61.8	61.4
ja	55.3	59.5	58.7	63.8	54.8	52.8	56.9	57.6	55.4	50.2	55.5	56.3	58.7	56.1	53.0	53.9	55.3	55.7	59.1	–	55.0	58.2
ko	53.5	61.7	55.0	49.2	58.0	54.7	49.3	58.2	55.6	59.0	61.6	48.3	53.5	60.2	54.3	58.7	64.4	64.1	52.3	58.3	–	56.5
zh	33.3	45.2	39.0	40.2	37.6	39.0	41.9	38.0	42.5	40.8	38.7	35.8	39.3	41.7	42.4	41.4	42.6	52.9	54.0	40.2	46.5	–

Table C27: Direction-level translation quality for DeepSeek Flash. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	52.8	50.7	55.9	55.2	57.1	47.5	50.8	48.6	48.6	57.9	54.5	59.1	53.2	61.7	64.0	55.8	59.4	57.4	46.2	48.1	54.1
cs	72.6	–	69.6	78.1	81.1	80.3	70.6	84.3	76.0	73.7	75.7	73.1	78.5	77.7	80.9	73.2	85.1	79.6	76.8	71.7	75.8	76.6
de	46.6	57.6	–	58.7	56.4	54.4	60.7	50.3	57.1	66.9	58.5	52.8	60.7	61.8	59.1	58.0	54.3	61.5	68.3	54.3	61.7	58.6
nl	71.5	70.1	67.0	–	65.0	88.8	67.5	69.7	72.6	75.7	69.2	63.5	79.5	75.2	66.1	71.7	83.4	84.2	78.7	76.7	68.2	72.8
fr	68.6	70.1	63.8	72.4	–	65.5	61.0	64.0	71.7	65.0	71.0	68.5	64.0	75.0	67.1	72.5	70.7	76.8	72.4	64.0	66.6	63.9
it	52.8	61.1	52.8	57.2	58.9	–	56.5	44.9	55.7	60.7	55.0	57.4	57.5	46.7	47.6	52.9	50.5	62.7	55.5	58.5	49.7	51.9
es	64.1	74.2	70.7	67.8	69.7	70.1	–	73.7	68.0	64.7	66.7	71.1	68.2	75.3	74.4	69.0	68.7	77.6	73.0	70.9	68.9	72.5
pt	45.9	57.6	58.8	67.0	58.6	56.7	58.3	–	56.5	56.1	65.6	66.7	66.3	59.2	57.1	60.7	61.2	74.5	68.9	65.1	65.0	56.1
tr	62.9	61.5	66.2	62.8	63.2	64.8	59.9	64.8	–	68.2	65.3	66.6	70.2	63.3	62.3	64.2	60.0	71.0	69.1	59.3	64.8	62.7
id	75.1	71.5	68.4	74.1	69.2	73.9	69.4	80.1	75.5	–	65.7	77.0	70.0	77.8	66.1	84.4	79.8	69.3	77.7	67.2	80.8	70.4
ms	60.9	68.5	60.4	62.1	55.5	62.9	58.2	48.4	60.8	50.8	–	64.2	64.3	65.9	67.9	74.3	72.0	67.0	63.8	64.9	71.4	58.2
vi	57.3	69.5	65.8	62.0	70.3	65.7	65.0	64.5	63.5	62.3	67.5	–	68.7	63.9	62.2	61.7	72.1	76.7	63.7	65.9	64.9	69.4
ru	59.7	67.7	65.6	67.1	54.3	66.9	64.7	59.6	66.1	64.6	64.1	67.8	–	60.8	57.5	63.5	63.0	77.0	71.1	63.2	65.6	61.5
uk	56.0	47.1	55.4	59.0	53.2	51.9	54.3	55.1	60.4	61.0	59.8	58.3	56.5	–	69.1	72.7	55.3	72.2	58.6	49.8	69.8	61.0
ar	49.7	52.5	55.4	46.8	53.9	50.5	54.3	46.5	53.7	52.6	55.1	53.8	52.5	48.3	–	46.7	53.4	60.1	56.5	54.1	48.0	52.8
he	48.0	51.3	46.8	47.6	47.7	46.9	45.6	53.1	47.6	54.6	51.2	50.2	44.1	47.0	49.9	–	54.2	60.7	47.5	45.6	52.8	43.8
hi	64.7	70.3	65.7	68.3	62.3	62.4	60.8	72.6	64.1	67.2	65.1	71.7	61.6	70.7	67.7	66.6	–	75.5	61.0	59.3	65.9	60.2
my	39.5	40.8	30.8	40.0	30.0	36.4	33.9	32.0	34.4	32.7	36.3	37.0	31.1	32.0	31.6	37.1	34.9	–	34.0	36.1	33.0	36.0
th	48.9	59.6	57.8	56.4	50.3	45.8	49.6	52.9	57.7	51.5	51.7	54.1	55.9	51.3	58.8	57.0	52.9	59.7	–	51.2	55.1	54.7
ja	59.8	62.0	58.2	60.8	65.7	56.1	62.7	58.2	54.6	55.0	58.5	54.6	61.0	62.3	51.2	63.4	64.2	64.9	62.3	–	55.3	60.5
ko	55.2	60.4	62.4	53.3	65.3	59.2	60.9	55.2	61.7	61.4	67.7	51.1	52.7	61.5	54.7	58.5	63.7	61.4	62.2	58.1	–	59.5
zh	61.8	64.6	61.0	61.3	56.8	61.8	57.6	62.0	64.0	59.8	59.3	61.2	63.6	62.1	59.8	65.5	66.5	66.6	65.2	55.2	61.5	–

Table C28: Direction-level translation quality for Qwen3.8 27B. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	63.1	60.9	56.9	57.8	65.2	55.0	57.8	52.1	66.8	62.6	61.7	60.5	64.8	62.0	65.3	59.3	62.1	62.2	67.8	60.8	60.4
cs	67.7	–	61.8	64.6	64.4	65.8	61.9	55.5	58.9	65.5	69.8	60.3	61.5	65.9	73.0	61.0	68.3	71.5	61.7	68.7	54.4	63.7
de	56.3	53.6	–	63.0	56.4	55.0	59.9	60.7	61.5	65.6	57.2	57.0	56.0	62.0	59.5	53.3	64.8	62.2	63.5	57.7	54.7	59.3
nl	67.2	68.2	59.0	–	67.3	64.8	77.3	68.1	70.1	68.5	71.2	70.7	72.9	67.8	65.3	78.3	74.8	79.4	84.2	85.1	73.4	74.7
fr	62.9	67.4	63.5	57.0	–	55.9	61.6	52.3	65.1	64.7	71.4	61.0	65.2	60.7	66.7	60.4	67.0	73.8	73.2	62.8	62.3	60.6
it	59.3	60.5	62.9	65.2	62.6	–	57.5	59.9	69.1	57.9	58.6	59.4	65.1	53.3	56.8	58.1	56.4	70.4	62.3	62.4	56.5	55.5
es	65.4	63.3	69.0	60.0	65.4	69.9	–	63.3	64.0	63.1	64.9	66.3	69.3	67.8	70.6	58.5	75.1	68.6	67.8	71.9	67.0	68.5
pt	58.0	64.0	56.6	65.3	61.2	69.7	71.2	–	65.7	64.2	62.5	68.4	59.7	66.8	65.4	50.1	69.3	66.0	67.9	59.0	55.2	67.4
tr	64.4	66.8	61.7	65.4	62.8	59.4	57.8	70.4	–	67.2	57.0	65.6	64.3	66.9	57.5	59.9	64.5	58.8	50.6	58.7	60.9	61.2
id	64.7	64.8	60.8	65.5	57.2	61.7	58.6	70.4	65.5	–	63.6	75.5	67.1	69.6	65.9	61.9	68.8	62.9	68.7	61.3	69.8	67.4
ms	62.1	63.4	69.0	60.7	64.7	64.8	60.6	60.5	61.7	69.7	–	67.3	64.3	60.5	62.0	64.2	59.0	50.4	71.0	67.4	64.2	60.3
vi	63.5	67.6	64.6	58.8	60.5	62.0	65.6	65.7	61.7	60.7	61.4	–	57.6	60.3	57.1	53.8	66.8	65.4	68.8	69.6	62.7	62.9
ru	66.4	73.4	74.8	70.0	68.6	77.7	71.3	71.9	74.8	70.2	65.1	70.4	–	73.1	72.3	67.2	71.8	69.8	83.3	77.5	72.9	69.2
uk	57.6	61.9	53.9	60.6	48.2	62.7	67.3	65.8	56.9	63.5	65.8	62.7	60.5	–	66.8	60.9	70.7	67.0	63.9	61.4	68.4	61.1
ar	56.7	56.5	63.6	57.5	61.0	51.6	62.2	52.4	61.9	52.7	59.3	59.6	59.6	53.3	–	62.4	61.2	58.8	58.2	61.3	56.6	60.8
he	50.0	50.8	54.8	52.5	54.0	53.7	51.9	54.7	51.2	59.4	56.5	55.5	46.1	54.9	51.4	–	58.7	55.7	56.5	55.8	58.5	54.1
hi	58.3	61.0	62.1	57.1	59.6	62.4	58.2	64.5	64.4	57.8	66.7	63.3	59.7	58.3	59.7	63.2	–	66.7	57.5	57.3	53.7	56.3
my	22.8	30.0	28.5	24.4	25.9	22.7	22.5	26.7	26.9	23.7	22.3	23.6	24.7	19.1	29.0	26.8	28.3	–	28.0	24.6	26.1	25.8
th	45.9	54.0	58.8	52.7	47.6	48.2	47.3	51.3	52.4	53.3	53.8	49.4	51.0	47.4	52.6	47.9	61.5	60.5	–	56.6	57.9	50.8
ja	68.2	57.8	71.6	67.2	65.1	59.1	61.1	63.8	65.1	59.7	67.7	63.3	65.4	66.6	61.7	66.4	63.2	59.1	69.4	–	66.3	65.5
ko	69.3	63.3	62.5	59.9	68.1	61.4	55.5	63.8	61.8	59.6	64.0	59.1	72.3	68.7	58.7	63.0	63.7	65.4	65.2	60.2	–	63.9
zh	65.1	62.0	62.8	66.1	65.1	68.2	65.4	61.6	68.1	66.2	66.0	66.0	68.6	60.7	65.7	61.9	63.4	67.2	67.0	62.3	61.2	–

Table C29: Direction-level translation quality for Claude Sonnet 5. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	65.4	61.3	57.6	54.5	61.9	52.7	57.9	56.3	56.5	60.7	68.8	63.9	68.0	63.6	64.0	58.2	72.3	63.8	51.8	58.4	58.5
cs	83.6	–	84.3	80.1	74.0	81.2	78.5	82.5	86.8	76.6	85.1	75.7	81.4	80.9	83.0	80.5	78.2	83.4	81.8	85.1	79.8	78.4
de	57.1	65.3	–	61.1	65.7	63.7	61.3	57.2	61.0	67.9	59.7	62.1	64.1	69.3	55.6	65.1	68.8	76.0	62.1	65.0	62.8	54.9
nl	79.4	84.9	84.2	–	93.0	94.0	84.3	79.9	92.7	87.7	90.5	92.7	91.2	88.9	85.1	90.9	89.5	85.9	87.5	86.1	82.6	83.4
fr	69.3	76.5	75.4	66.7	–	67.7	76.3	74.6	69.3	67.0	81.3	74.2	67.6	67.4	75.2	69.9	79.8	79.4	77.8	75.3	75.1	64.6
it	63.2	67.5	63.7	69.1	68.3	–	67.5	58.8	60.5	68.8	70.3	68.1	66.8	65.9	65.4	67.2	66.8	75.5	69.2	68.3	71.2	60.0
es	66.9	64.5	64.6	66.4	60.7	71.5	–	66.8	65.4	67.5	67.8	68.7	68.7	69.6	76.0	71.9	72.9	73.8	71.7	60.2	66.1	70.7
pt	58.5	64.3	56.7	66.8	66.5	69.0	65.8	–	64.8	67.7	70.1	70.7	62.5	70.5	67.1	68.1	62.4	79.6	65.6	56.3	66.5	68.4
tr	67.8	67.7	65.0	65.4	72.5	66.8	67.7	65.9	–	70.2	62.8	74.8	66.6	70.6	64.4	64.5	65.2	70.4	68.0	65.2	67.3	59.5
id	79.6	76.0	73.8	76.1	68.8	73.6	76.4	80.1	73.7	–	77.7	81.2	77.5	76.6	73.1	84.9	81.9	77.9	79.9	73.4	83.2	76.5
ms	71.0	76.3	68.9	73.0	63.2	66.2	73.8	70.2	75.8	71.3	–	79.7	66.7	69.8	76.4	73.2	75.4	86.9	72.8	67.6	75.8	70.5
vi	71.6	73.6	68.8	67.4	72.8	73.6	81.1	74.8	78.1	72.3	79.7	–	79.8	75.7	74.5	70.5	87.1	76.8	74.4	79.3	73.7	71.6
ru	70.4	69.3	70.6	79.1	72.5	75.4	77.2	70.4	76.6	73.9	77.6	71.5	–	79.6	76.5	78.5	75.0	81.6	76.8	77.3	73.9	76.1
uk	67.3	70.3	72.6	73.4	62.2	69.8	67.1	70.1	69.9	77.5	75.2	73.4	66.1	–	67.8	68.7	75.9	73.8	74.9	65.3	79.1	73.2
ar	62.2	63.2	69.2	68.0	61.3	67.6	69.5	61.6	70.9	66.7	68.2	66.1	69.3	66.3	–	69.3	68.4	67.9	65.8	65.2	68.2	65.9
he	67.0	64.0	65.5	65.9	69.3	65.0	66.6	70.3	67.4	73.8	64.8	63.8	64.4	63.2	67.2	–	73.6	73.9	65.4	71.3	66.4	67.7
hi	75.2	73.6	74.8	74.3	70.1	75.1	77.2	80.8	80.4	69.4	76.0	80.7	73.6	81.4	79.2	82.4	–	76.1	75.4	80.3	77.9	74.2
my	39.2	45.9	42.3	46.3	42.2	42.7	37.4	39.0	41.7	40.6	48.5	47.4	36.8	46.9	40.5	50.1	43.8	–	45.9	46.1	49.6	45.7
th	68.7	67.8	70.5	70.3	63.0	68.2	64.1	70.8	67.7	66.2	63.7	71.5	69.7	64.8	73.4	72.8	72.9	69.2	–	67.0	71.3	66.3
ja	65.6	67.3	69.7	64.2	67.6	65.3	65.1	67.1	67.5	66.6	65.0	66.6	68.0	62.1	64.9	72.5	70.0	70.7	65.8	–	66.2	65.9
ko	71.9	69.9	73.0	73.1	72.4	73.0	70.8	68.7	75.6	75.7	75.1	71.1	76.5	73.6	75.9	70.3	72.4	77.3	66.5	72.5	–	72.1
zh	59.9	63.4	65.0	68.5	60.9	62.9	60.9	63.2	68.4	64.7	65.8	67.9	65.5	65.1	65.0	66.7	64.9	74.4	74.0	64.8	68.4	–

Table C30: Direction-level translation quality for Muse Spark 1.3. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	64.5	59.5	61.0	61.8	65.5	50.0	62.4	63.7	63.5	64.2	58.3	65.1	70.6	68.6	68.4	66.4	75.3	68.6	60.2	64.2	58.9
cs	82.7	–	80.7	81.2	88.3	80.8	85.4	89.8	95.9	81.7	89.0	82.4	87.2	94.8	83.5	83.2	86.4	84.3	88.4	79.8	75.5	87.5
de	46.3	61.3	–	62.4	64.4	62.5	56.2	58.1	67.2	67.4	60.2	62.2	62.9	70.4	63.0	63.1	61.6	70.2	63.9	56.0	58.9	62.9
nl	81.4	79.8	75.1	–	83.3	81.9	86.6	85.0	85.3	77.4	80.8	84.9	82.5	77.3	87.5	85.9	76.3	86.1	79.1	83.9	81.8	83.7
fr	64.4	86.3	73.1	75.7	–	76.7	78.0	71.9	81.9	76.5	78.1	64.8	82.5	79.4	77.1	77.4	81.7	83.4	79.7	76.5	72.0	71.7
it	60.9	72.1	68.1	70.3	68.4	–	65.8	60.9	64.0	66.9	68.6	65.8	63.8	70.7	69.6	71.3	64.8	67.0	67.2	76.2	62.6	64.2
es	69.0	80.2	75.8	73.5	79.5	71.6	–	73.4	75.1	77.1	74.2	72.8	71.9	75.8	67.7	73.0	71.3	76.9	71.1	71.1	73.3	76.6
pt	62.2	62.6	66.2	74.5	64.8	71.1	58.0	–	65.6	68.5	68.6	70.5	65.7	72.2	72.3	62.3	60.9	81.5	68.6	66.4	70.2	66.5
tr	71.4	79.5	74.3	73.8	74.9	75.7	71.8	70.8	–	72.0	71.6	67.5	73.9	78.4	79.6	71.7	65.5	76.0	70.5	76.8	68.8	73.2
id	75.0	82.0	79.7	73.3	78.2	72.3	66.6	77.9	79.8	–	71.8	74.1	72.2	75.2	73.1	77.3	81.5	81.7	74.5	79.4	81.0	74.7
ms	67.2	81.5	72.9	77.2	73.2	72.9	66.7	70.5	75.4	76.3	–	74.6	67.6	70.2	69.1	81.0	81.8	84.3	78.4	70.8	77.6	75.1
vi	83.7	82.9	84.8	75.7	83.7	85.2	85.4	82.0	81.7	80.3	83.5	–	80.4	81.1	75.0	83.3	85.7	80.5	77.7	80.8	79.6	81.2
ru	79.4	76.8	77.1	78.2	76.0	83.7	80.9	82.2	86.2	79.0	74.4	72.8	–	80.1	80.5	72.9	74.2	81.4	86.1	78.8	81.1	82.4
uk	63.5	72.5	69.9	79.6	66.5	68.2	63.6	65.8	70.3	70.0	70.9	70.4	66.3	–	74.6	79.0	71.7	74.5	71.5	66.7	79.4	65.7
ar	68.9	73.2	66.2	74.4	71.7	67.9	69.3	69.3	74.0	70.2	69.9	64.3	73.2	63.6	–	70.4	71.8	74.2	74.1	74.2	63.2	69.8
he	73.9	72.8	72.7	69.0	72.3	71.7	74.5	75.1	68.5	75.5	76.6	68.7	69.0	68.5	73.6	–	76.9	76.4	72.2	70.7	73.7	74.9
hi	75.9	78.3	80.1	75.4	75.3	80.9	79.7	79.9	76.8	79.0	79.8	81.6	74.2	80.9	82.8	76.8	–	80.0	74.5	78.9	77.3	78.0
my	45.3	50.3	38.9	42.4	41.7	44.3	44.7	46.3	43.5	37.3	42.1	45.0	43.0	46.1	42.8	43.8	45.7	–	39.7	42.5	42.1	43.8
th	66.3	73.7	70.9	71.5	69.5	71.9	67.7	74.9	77.1	72.7	72.7	70.9	70.7	73.6	79.1	72.5	76.0	73.8	–	68.2	72.0	72.7
ja	71.0	69.5	71.0	72.8	71.9	70.5	69.2	67.4	68.1	65.3	69.3	62.9	70.0	68.3	65.3	72.4	67.9	74.2	68.4	–	69.8	67.0
ko	75.8	75.7	71.7	71.0	69.8	79.7	72.9	71.3	79.0	74.2	76.2	70.7	74.4	75.9	74.0	72.5	76.0	81.2	69.6	75.9	–	76.9
zh	60.5	71.2	63.7	67.9	64.2	67.8	60.7	65.8	70.8	61.3	68.2	62.4	66.1	62.6	62.8	66.5	67.8	70.9	73.5	64.4	68.7	–

Table C31: Direction-level translation quality for Qwen3.8 Max. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	74.9	69.4	62.2	64.5	70.2	61.6	60.7	64.7	66.0	73.0	67.4	73.3	63.9	69.3	76.2	68.2	63.5	73.3	79.6	64.7	65.6
cs	85.4	–	76.7	80.2	85.2	82.6	77.2	88.6	86.1	83.0	91.9	85.2	85.5	87.0	89.8	88.5	87.2	75.2	86.9	82.6	80.9	88.9
de	60.8	67.9	–	72.3	77.2	68.9	63.5	65.0	69.5	72.0	71.2	70.3	65.2	75.1	72.0	63.7	69.9	68.3	70.6	70.3	70.2	70.4
nl	83.2	83.7	80.5	–	83.6	77.0	88.2	81.8	80.5	80.7	85.1	85.1	89.8	82.6	80.3	82.9	85.0	70.3	78.7	86.3	78.4	77.9
fr	75.1	77.2	74.7	74.5	–	65.8	74.0	70.4	79.9	72.9	78.2	74.9	72.8	75.1	77.2	78.7	79.6	64.3	74.9	75.4	73.1	66.7
it	72.6	72.3	74.2	75.4	69.0	–	69.0	65.4	76.4	70.5	68.7	73.4	72.3	74.2	72.8	73.3	70.1	66.0	78.5	77.0	71.7	68.4
es	74.8	78.4	74.4	77.8	71.4	78.7	–	69.9	76.8	72.2	78.1	77.5	76.0	77.4	83.3	77.4	71.2	68.0	76.8	77.0	76.1	75.1
pt	72.5	73.4	76.6	76.8	74.6	72.3	72.7	–	76.4	73.6	75.0	78.7	79.3	68.5	71.4	74.3	72.6	70.1	77.9	69.7	74.6	69.4
tr	76.4	74.3	74.9	69.2	74.8	63.9	69.6	68.6	–	64.4	66.6	75.6	70.9	68.0	67.2	71.2	68.5	66.3	70.3	75.5	72.6	67.8
id	78.6	79.9	68.1	80.5	75.8	71.4	79.7	76.2	82.7	–	75.9	75.7	76.6	74.6	75.1	82.4	80.5	62.7	79.1	73.0	84.3	76.1
ms	69.6	75.9	65.4	73.1	69.8	76.5	64.3	71.4	78.0	71.8	–	72.3	73.2	64.4	73.9	84.4	78.3	75.9	78.0	72.5	69.6	71.7
vi	81.0	84.8	81.6	82.7	76.2	80.9	83.1	71.4	81.9	78.6	79.3	–	80.5	81.7	73.0	80.6	87.7	69.5	78.6	85.4	75.7	80.1
ru	77.4	76.0	77.2	80.9	77.3	78.9	78.0	72.7	79.2	79.2	77.2	77.3	–	81.9	76.8	77.0	81.1	63.9	78.0	73.1	77.3	78.4
uk	65.0	70.6	65.0	70.9	63.8	69.7	64.3	69.7	70.9	72.4	74.8	71.9	59.2	–	70.5	71.8	72.5	66.7	66.9	70.7	74.7	70.4
ar	69.6	67.6	66.4	74.0	71.9	70.1	70.2	65.2	70.4	70.6	69.1	66.2	68.1	67.0	–	68.3	69.7	65.9	65.2	72.1	66.1	68.9
he	65.6	68.1	64.5	61.5	66.1	62.8	66.6	65.2	66.2	70.2	71.5	69.5	66.3	67.5	65.1	–	73.2	66.7	70.6	72.6	71.0	69.6
hi	80.3	80.2	81.6	81.8	78.9	81.0	80.5	79.5	82.5	73.5	81.8	83.7	77.8	79.9	83.2	76.8	–	77.1	73.7	76.8	77.9	76.4
my	42.3	47.2	38.9	45.4	38.4	42.3	37.2	43.1	39.8	43.4	40.9	48.7	36.8	44.3	37.6	43.4	47.0	–	40.9	46.1	40.0	42.4
th	64.0	63.8	72.4	70.5	68.6	67.1	71.3	70.7	69.8	69.1	67.8	72.4	75.9	68.5	70.5	73.0	71.7	68.9	–	74.2	72.7	70.7
ja	74.1	65.9	70.8	75.8	76.8	70.8	77.4	68.1	70.9	70.3	76.0	77.6	75.3	73.5	72.8	71.1	76.7	67.3	73.4	–	71.2	72.0
ko	84.8	83.0	84.6	79.6	80.9	77.9	84.4	82.6	80.7	79.0	82.8	78.0	81.4	83.3	78.1	75.8	84.9	71.8	78.0	84.9	–	82.8
zh	68.3	73.1	67.0	72.7	72.4	75.4	70.8	70.8	78.3	73.6	73.4	73.5	73.9	74.9	72.0	76.8	76.5	60.9	78.8	69.4	74.2	–

Table C32: Direction-level translation quality for Gemini 3.8 Flash. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	69.4	60.0	68.1	67.0	71.1	57.5	65.0	61.8	66.1	67.9	65.8	64.0	67.4	64.7	78.4	61.5	73.7	74.8	67.7	68.6	63.3
cs	88.1	–	88.6	84.0	83.7	81.5	90.0	86.5	87.0	87.0	85.3	83.3	87.4	86.7	82.0	89.8	89.0	88.8	84.8	85.5	86.7	84.8
de	65.5	71.9	–	75.2	77.2	66.3	68.8	69.7	66.3	73.9	68.0	64.2	67.0	76.7	64.4	70.0	71.8	78.9	71.8	72.5	66.8	68.7
nl	88.6	93.8	93.5	–	93.5	89.4	87.0	87.8	86.7	97.8	93.8	87.9	98.0	89.7	88.2	83.3	85.8	86.0	81.6	80.4	94.4	82.2
fr	73.7	80.9	79.6	75.6	–	70.8	79.7	73.7	79.4	73.9	79.3	83.1	72.8	77.7	84.7	83.6	80.1	83.0	81.3	76.2	77.4	74.9
it	68.3	74.5	70.8	76.8	72.4	–	69.5	73.9	69.1	72.9	69.8	73.5	69.1	76.1	69.3	70.7	71.6	76.5	72.4	74.8	73.0	67.9
es	66.9	83.5	69.5	73.3	67.1	72.3	–	74.9	77.2	74.5	78.9	83.2	73.0	71.8	79.6	79.3	75.8	79.0	75.6	76.1	73.8	70.3
pt	76.7	75.3	71.3	77.5	80.0	86.0	71.8	–	74.2	79.7	75.2	73.7	72.8	76.5	71.8	74.9	74.2	84.2	67.9	72.2	70.7	69.6
tr	79.9	76.9	79.3	76.9	75.6	76.0	82.0	75.9	–	76.4	73.1	74.7	80.0	84.1	69.8	75.6	66.5	76.4	78.5	74.2	83.0	70.3
id	76.9	83.7	77.3	85.3	81.3	78.5	74.1	84.0	86.0	–	82.9	87.5	83.4	89.4	84.6	91.2	83.4	82.8	86.2	83.3	84.6	82.0
ms	76.3	78.4	79.0	73.8	74.8	70.1	74.2	77.4	72.7	75.7	–	77.2	73.0	77.5	69.8	80.3	81.6	84.1	74.5	71.3	80.8	75.1
vi	85.8	84.8	83.8	76.4	85.1	84.5	80.4	78.3	80.4	82.8	82.1	–	80.5	83.4	78.3	81.7	85.7	87.7	81.6	82.2	80.4	80.8
ru	74.0	73.8	76.1	73.0	76.7	74.8	69.1	71.0	82.1	72.4	74.2	73.8	–	78.6	76.7	78.3	74.0	75.8	88.7	75.3	74.8	74.1
uk	69.8	70.1	64.4	64.6	70.3	71.2	65.9	70.2	68.8	79.9	75.3	79.7	67.9	–	70.6	73.7	64.2	64.7	72.6	62.1	71.4	63.7
ar	67.3	70.8	71.0	67.9	70.8	69.9	71.0	65.1	70.3	67.4	76.8	71.5	76.1	66.9	–	72.6	71.2	70.8	73.4	72.1	68.4	67.4
he	69.9	71.2	68.7	66.9	72.0	71.4	77.2	75.2	69.7	73.8	70.4	71.9	64.9	70.3	69.4	–	74.4	72.8	70.7	72.4	74.3	72.4
hi	78.0	80.9	82.1	79.2	75.5	78.3	79.4	83.4	82.1	77.2	82.8	81.4	80.8	85.2	78.8	83.6	–	81.8	80.2	80.7	77.8	75.0
my	44.1	44.4	42.1	43.0	39.3	46.4	36.0	41.7	45.6	38.5	41.6	47.1	46.5	41.4	44.5	44.5	48.2	–	39.9	38.9	44.5	41.0
th	70.3	66.7	70.6	71.5	70.1	67.4	68.5	68.6	66.6	68.4	65.7	69.9	71.4	64.6	74.6	76.2	72.7	71.9	–	71.7	67.6	69.7
ja	74.4	68.5	76.8	72.7	75.5	74.9	69.8	74.7	71.0	70.6	72.5	76.3	69.7	68.3	69.9	75.0	72.2	77.1	74.0	–	74.4	71.5
ko	82.6	83.1	86.6	80.2	83.5	83.0	84.3	81.0	82.6	78.2	83.6	81.4	85.5	78.6	80.4	80.1	80.4	82.2	82.3	78.2	–	80.1
zh	70.0	74.1	70.6	75.8	72.1	70.6	68.1	74.8	73.0	72.8	70.3	70.3	69.7	71.8	71.5	71.6	73.4	75.1	79.0	68.6	73.2	–

Table C33: Direction-level translation quality for Doubao Seed 2.1 Pro. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	74.0	70.7	66.7	73.8	79.9	63.8	76.8	72.9	73.8	75.8	73.0	73.3	70.9	77.9	74.8	61.4	67.5	75.2	69.4	68.5	74.4
cs	90.9	–	82.8	85.6	92.3	84.0	90.3	81.8	92.2	89.1	86.5	85.5	94.2	88.4	84.4	81.7	83.9	92.9	88.2	86.7	79.4	91.4
de	64.9	74.1	–	67.9	68.8	68.3	66.0	62.7	64.3	74.2	70.6	61.6	77.1	76.5	65.9	74.6	74.2	71.1	73.6	64.6	67.1	70.8
nl	82.3	86.4	84.8	–	92.6	86.3	75.8	90.6	90.7	91.4	78.9	90.6	86.7	90.9	89.2	81.6	88.4	76.4	86.3	86.0	88.5	83.6
fr	75.3	88.2	75.0	80.8	–	80.9	81.8	80.2	73.6	78.9	79.4	79.9	82.5	75.8	84.1	82.7	82.7	91.0	81.7	73.4	85.8	77.3
it	71.3	73.6	76.4	76.2	68.9	–	68.0	62.5	73.8	73.1	64.5	72.6	68.1	68.0	73.5	72.8	68.1	70.2	75.6	71.2	74.3	66.9
es	74.6	86.1	76.1	78.1	77.3	79.2	–	74.2	82.8	81.1	81.9	77.8	82.0	85.0	86.0	70.9	77.6	78.6	77.0	79.5	78.5	76.1
pt	69.5	74.7	66.2	67.4	66.8	70.2	65.4	–	65.7	70.0	68.4	74.2	66.5	73.7	77.9	68.0	63.3	78.5	69.4	62.1	69.7	66.4
tr	84.2	74.8	82.1	73.1	83.3	78.6	77.3	73.2	–	78.1	77.4	78.2	78.0	79.8	78.0	83.2	69.3	76.0	78.7	79.4	73.4	73.2
id	82.1	80.3	69.2	78.2	79.5	84.9	75.7	83.2	79.2	–	77.9	82.2	84.5	79.9	81.5	82.7	83.9	84.1	86.1	74.3	77.6	74.6
ms	77.9	85.4	76.7	78.4	81.9	78.1	81.3	79.4	77.5	76.0	–	79.8	81.3	67.1	86.1	80.0	85.9	93.1	86.1	82.5	80.9	81.0
vi	76.6	84.8	84.4	78.8	84.1	79.6	87.1	82.4	83.6	81.5	81.3	–	78.2	78.9	81.3	81.1	82.5	84.6	84.6	87.0	75.6	81.6
ru	75.8	71.9	77.5	71.0	75.0	81.5	71.8	87.8	82.5	76.4	73.9	75.9	–	68.4	80.8	76.4	74.9	84.4	78.8	75.3	77.8	77.2
uk	67.2	73.5	68.3	77.8	64.7	73.7	69.5	67.0	73.1	65.6	71.7	80.4	63.9	–	77.0	71.4	71.1	73.8	66.6	64.1	70.6	66.7
ar	69.6	66.5	70.6	74.4	71.8	73.5	73.9	68.7	70.0	65.3	73.8	72.5	74.7	68.3	–	72.7	69.2	72.7	70.9	71.7	68.5	74.8
he	66.6	68.9	62.6	62.5	63.0	70.2	66.5	67.0	67.6	70.9	72.0	67.5	64.3	62.0	70.3	–	68.4	76.1	66.2	65.8	71.3	67.5
hi	79.5	83.8	85.2	83.8	78.2	84.3	78.7	83.7	81.6	82.9	82.3	85.1	81.3	85.1	80.3	84.4	–	82.4	80.0	83.4	82.0	80.7
my	44.4	55.8	41.2	44.8	46.1	47.7	41.6	41.8	46.0	42.8	40.7	52.7	44.2	46.0	45.2	50.5	50.0	–	40.0	47.2	50.2	46.9
th	65.5	72.6	71.2	70.1	68.0	65.2	71.8	73.4	70.1	70.2	71.8	73.6	73.7	73.2	77.6	73.3	72.4	71.5	–	69.1	76.2	71.0
ja	74.3	74.8	71.4	78.1	75.1	74.2	73.0	70.6	72.6	73.4	73.4	71.1	79.8	74.8	74.2	75.6	79.6	73.8	82.6	–	72.0	77.9
ko	80.6	83.4	85.6	78.1	81.4	85.2	82.8	83.6	78.8	79.2	80.5	80.8	81.4	84.6	79.0	87.0	79.6	89.4	76.5	87.0	–	82.5
zh	72.8	74.6	74.1	75.4	76.5	75.5	67.3	73.3	77.0	75.4	77.8	79.3	79.5	72.3	76.7	73.1	76.8	77.7	77.7	70.2	75.2	–

Table C34: Direction-level translation quality for GPT-6 Astra. Rows are source languages and columns are target languages.

Src\Tgt	en	cs	de	nl	fr	it	es	pt	tr	id	ms	vi	ru	uk	ar	he	hi	my	th	ja	ko	zh
en	–	81.9	79.3	78.5	70.3	77.3	82.6	76.4	83.6	76.9	78.1	80.1	79.6	77.9	79.3	79.7	75.1	84.7	82.3	78.4	76.7	78.5
cs	90.9	–	82.7	92.0	92.1	86.5	91.5	89.5	87.8	92.0	87.8	87.6	83.4	94.0	87.6	90.6	88.0	89.1	87.4	84.6	93.1	88.3
de	73.9	75.0	–	75.9	73.1	68.8	74.2	62.2	79.9	76.8	72.1	72.4	77.2	83.2	73.3	69.3	72.6	85.1	75.3	78.9	71.2	74.1
nl	86.4	83.9	88.6	–	85.8	87.0	79.6	91.8	89.9	90.7	84.2	86.2	84.3	80.8	88.5	85.5	85.8	90.7	92.1	86.5	81.9	89.4
fr	76.6	85.6	79.0	80.1	–	74.2	79.0	88.8	88.7	80.2	82.8	82.9	80.1	80.0	82.0	84.4	84.5	85.9	85.1	82.5	77.7	80.0
it	69.1	76.0	77.2	79.4	79.6	–	76.7	69.1	74.3	74.5	76.6	76.4	78.7	82.5	73.8	76.0	78.1	79.7	76.1	79.9	74.5	76.2
es	78.0	84.1	81.0	82.0	77.7	78.6	–	80.8	78.6	80.5	84.2	86.1	82.1	81.9	88.3	91.1	77.5	87.9	77.8	79.2	82.7	79.9
pt	73.0	72.8	68.1	74.3	75.3	77.2	72.8	–	79.8	80.1	73.4	77.1	77.4	69.1	77.5	77.6	79.6	84.8	83.1	73.6	73.2	78.4
tr	74.8	79.2	81.1	83.6	85.0	78.0	79.8	75.6	–	79.5	75.1	85.5	79.1	80.5	73.6	81.3	76.6	73.3	79.4	84.5	76.3	78.0
id	73.0	79.0	77.6	72.8	77.7	74.4	68.2	85.5	78.4	–	75.3	78.7	81.5	84.0	72.9	81.0	88.5	83.8	85.7	82.1	88.1	80.0
ms	72.9	84.5	81.6	78.8	78.7	78.4	81.9	79.0	79.0	78.9	–	84.6	83.7	79.2	81.0	79.7	87.4	89.4	76.5	79.8	84.4	81.3
vi	83.5	83.9	86.5	82.4	91.9	84.9	86.5	83.3	86.8	85.5	86.5	–	86.1	87.7	81.4	85.7	80.0	86.4	89.7	86.0	76.4	85.6
ru	72.1	77.2	77.2	81.3	78.3	83.7	78.2	82.7	79.3	78.1	81.3	79.3	–	79.2	79.0	70.8	75.2	76.7	80.1	86.4	81.3	84.5
uk	67.9	69.0	74.0	76.3	67.3	75.1	67.8	66.1	70.9	73.4	78.8	74.1	71.5	–	76.1	77.2	74.4	80.1	75.1	73.6	75.4	68.9
ar	71.4	67.8	69.5	69.7	73.0	73.1	71.1	72.2	72.9	71.9	75.0	68.7	65.6	66.9	–	71.5	72.0	70.2	72.8	72.1	73.0	68.0
he	73.8	72.1	74.9	71.1	74.2	77.3	77.2	78.8	72.1	76.2	76.3	70.1	68.2	71.9	74.4	–	78.0	75.1	76.6	78.6	74.7	76.6
hi	84.1	86.2	83.2	82.3	80.1	85.0	83.9	83.4	85.4	83.4	84.6	88.1	81.7	85.0	83.2	80.3	–	86.0	77.7	86.9	82.9	78.2
my	50.6	56.6	48.1	46.8	46.0	51.9	48.2	54.8	53.9	48.1	46.2	51.6	53.5	52.4	48.7	54.6	57.8	–	49.4	52.5	52.3	51.8
th	74.8	75.2	77.7	78.7	80.0	70.3	73.0	78.1	77.4	79.1	73.2	76.2	82.3	72.2	78.4	76.9	75.2	76.5	–	76.7	76.5	76.1
ja	76.4	77.8	82.9	80.8	80.5	80.4	76.4	76.7	79.1	75.8	82.3	75.5	80.7	80.5	76.4	81.1	83.1	78.7	79.8	–	80.4	78.6
ko	81.8	84.6	80.0	79.3	78.3	80.0	77.5	79.0	80.1	80.7	82.3	74.8	85.4	78.3	79.1	79.7	81.6	80.2	78.7	81.6	–	80.6
zh	71.1	76.4	70.7	77.1	73.8	76.4	72.7	73.7	74.6	75.5	71.2	67.5	74.0	74.8	75.9	72.4	76.0	80.4	75.8	69.6	71.6	–

C.3 Dimension Profiles by Source Language

Tables C35–C40 report T, K, and V separately for each model and source-image language, averaged over valid target languages. Each table presents two models. Blue and purple cells mark the best scores across the full closed- and open-weight groups, respectively, for the same language and metric.

Table C35: Source-language T, K, and V for GPT-5.6 Luna and Claude Sonnet 5.

	GPT-5.6 Luna			Claude Sonnet 5		
Source language	T	K	V	T	K	V
English	53.73	82.94	59.80	60.77	86.78	57.40
Czech	67.96	76.06	74.83	80.99	87.00	78.48
German	53.29	89.32	58.94	63.14	84.02	65.22
Dutch	66.38	72.53	71.70	87.33	87.40	76.93
French	66.59	89.26	65.32	72.89	90.41	58.43
Italian	59.00	73.63	54.02	66.78	81.52	56.65
Spanish	67.18	92.56	65.11	68.21	93.63	63.90
Portuguese	63.21	96.71	71.70	66.11	97.47	69.36
Turkish	61.18	92.20	59.26	67.06	83.24	51.73
Indonesian	60.93	94.16	72.42	77.24	95.67	77.27
Malay	65.21	98.55	75.75	72.59	98.56	77.22
Vietnamese	65.20	91.20	70.91	75.10	89.88	71.86
Russian	59.33	82.57	61.42	75.23	86.07	67.87
Ukrainian	50.22	90.45	48.96	71.12	98.21	75.95
Arabic	49.50	81.80	56.18	66.71	80.35	71.70
Hebrew	37.87	58.73	42.21	67.46	68.21	72.77
Hindi	52.52	53.86	59.04	76.58	93.06	73.25
Burmese	29.00	33.90	37.28	43.75	39.77	53.53
Thai	62.73	83.82	69.59	68.57	83.70	72.66
Japanese	56.27	63.32	48.71	66.85	81.09	57.03
Korean	56.50	56.99	52.28	72.74	76.21	71.08
Chinese	41.58	61.38	51.17	65.73	79.00	68.19

Table C36: Source-language T, K, and V for Muse Spark 1.3 and Qwen3.8 Max.

	Muse Spark 1.3			Qwen3.8 Max		
Source language	T	K	V	T	K	V
English	63.84	81.83	64.79	68.20	84.55	74.87
Czech	85.17	94.60	85.71	84.50	92.85	87.79
German	61.96	84.86	61.95	69.25	93.89	75.31
Dutch	82.18	86.92	81.34	81.98	85.79	80.35
French	76.60	94.72	66.25	74.07	93.83	73.39
Italian	67.10	80.29	60.20	71.95	96.26	71.67
Spanish	73.86	97.94	68.03	75.62	92.88	72.47
Portuguese	67.58	95.69	82.53	73.83	96.49	86.19
Turkish	73.23	86.46	66.37	70.32	93.67	60.90
Indonesian	76.25	95.05	81.41	76.63	88.59	78.22
Malay	74.48	99.12	87.54	72.85	96.27	78.79
Vietnamese	81.63	97.05	78.57	79.74	94.03	78.33
Russian	79.25	86.74	74.69	77.08	85.10	73.40
Ukrainian	70.51	98.50	71.20	69.16	99.75	72.90
Arabic	70.18	89.71	77.36	68.70	84.95	77.96
Hebrew	72.73	77.10	77.23	67.64	69.36	70.32
Hindi	78.39	91.26	75.54	79.28	88.19	81.18
Burmese	43.40	39.87	54.86	42.19	39.68	57.47
Thai	72.30	86.77	78.08	70.16	86.22	78.73
Japanese	69.14	88.56	61.35	72.74	84.80	64.71
Korean	74.50	87.95	73.53	80.93	92.42	81.70
Chinese	66.09	72.64	65.69	72.71	82.98	78.02

Table C37: Source-language T, K, and V for Gemini 3.8 Flash and Doubao Seed 2.1 Pro.

	Gemini 3.8 Flash			Doubao Seed 2.1 Pro		
Source language	T	K	V	T	K	V
English	66.86	83.81	70.67	72.12	88.55	81.98
Czech	86.21	96.00	86.54	87.25	96.30	89.02
German	70.28	86.24	75.82	69.47	96.84	83.53
Dutch	89.02	85.83	78.72	86.09	92.94	83.57
French	78.16	90.46	75.94	80.53	93.75	82.36
Italian	72.04	93.67	71.29	70.93	80.14	76.02
Spanish	75.03	96.51	68.66	79.07	96.62	80.97
Portuguese	75.06	98.13	84.75	69.24	97.34	80.50
Turkish	76.43	94.95	71.74	77.59	94.87	81.79
Indonesian	83.26	97.27	81.84	80.08	94.12	91.42
Malay	76.07	97.70	86.58	80.78	98.91	93.08
Vietnamese	82.22	96.77	78.56	81.89	97.95	85.02
Russian	75.59	85.58	70.80	76.90	86.31	76.98
Ukrainian	69.57	100.00	72.51	70.36	99.35	73.80
Arabic	70.41	87.65	74.93	71.14	92.23	79.64
Hebrew	71.42	77.88	77.68	67.48	69.79	77.40
Hindi	80.20	91.30	75.26	82.32	91.64	84.90
Burmese	42.83	39.66	54.40	45.98	41.13	63.92
Thai	69.74	87.18	79.16	71.51	86.87	84.21
Japanese	72.85	90.00	66.13	74.86	88.02	73.41
Korean	81.80	96.00	82.70	82.24	94.14	85.86
Chinese	72.22	81.48	73.55	75.15	90.33	81.85

Table C38: Source-language T, K, and V for GPT-6 Astra and Gemma-4 E4B.

	GPT-6 Astra			Gemma-4 E4B		
Source language	T	K	V	T	K	V
English	78.92	88.64	91.33	21.47	41.16	16.88
Czech	88.88	98.60	91.76	21.34	29.24	15.02
German	74.48	93.65	87.49	17.74	27.28	11.96
Dutch	86.64	82.22	80.39	17.90	48.85	24.21
French	81.91	92.90	86.47	26.43	28.76	23.43
Italian	76.41	95.90	81.22	18.99	29.29	14.33
Spanish	81.89	99.76	81.98	23.05	34.70	18.97
Portuguese	76.10	97.50	90.58	20.46	53.06	18.12
Turkish	79.04	94.39	79.25	28.63	46.62	22.42
Indonesian	79.44	93.08	91.59	25.55	33.52	26.16
Malay	81.00	96.90	87.61	25.61	54.98	20.62
Vietnamese	85.07	98.40	88.57	17.79	29.27	12.84
Russian	79.14	86.47	84.72	19.01	28.73	16.38
Ukrainian	73.00	99.14	79.58	22.54	52.75	24.84
Arabic	70.87	86.23	79.15	21.45	32.34	22.18
Hebrew	74.68	77.85	82.25	20.05	26.89	19.51
Hindi	83.41	89.57	80.85	36.94	51.81	35.66
Burmese	51.22	51.29	66.22	8.36	5.83	9.16
Thai	76.42	88.01	86.97	15.50	15.87	15.43
Japanese	79.23	87.74	82.01	10.28	15.53	9.44
Korean	80.18	86.62	85.16	13.40	16.06	11.37
Chinese	73.88	86.48	77.48	9.15	17.45	10.54

Table C39: Source-language T, K, and V for GLM-4.1V 9B and GLM-5.3 Flash.

	GLM-4.1V 9B			GLM-5.3 Flash		
Source language	T	K	V	T	K	V
English	35.74	63.93	39.42	56.24	82.04	59.21
Czech	32.92	40.99	47.09	65.22	70.29	70.43
German	36.66	58.27	40.70	53.03	84.82	57.28
Dutch	44.58	78.07	58.60	78.22	96.82	66.36
French	43.43	74.67	39.38	65.15	88.42	56.96
Italian	34.54	45.59	39.07	53.76	77.10	57.17
Spanish	38.50	62.46	42.42	60.56	91.95	52.87
Portuguese	38.54	85.60	49.51	63.83	95.48	66.09
Turkish	36.86	65.43	36.84	58.73	86.11	48.76
Indonesian	37.73	76.34	55.94	66.33	93.63	69.00
Malay	43.70	77.27	52.60	64.05	93.59	66.90
Vietnamese	29.55	50.00	35.32	60.63	84.28	64.58
Russian	38.21	51.06	37.83	64.50	84.04	62.00
Ukrainian	24.23	63.07	37.52	59.79	95.14	63.09
Arabic	14.60	14.44	14.41	38.56	55.65	49.31
Hebrew	6.25	9.82	8.89	35.11	51.89	46.18
Hindi	9.30	8.39	12.51	48.61	74.04	55.57
Burmese	15.13	12.20	16.60	22.90	12.33	30.78
Thai	11.67	14.35	14.17	33.46	41.51	43.39
Japanese	32.46	38.33	31.79	58.31	73.49	52.95
Korean	15.76	20.43	20.23	48.84	49.55	52.53
Chinese	33.79	45.55	38.34	58.82	75.22	62.23

Table C40: Source-language T, K, and V for DeepSeek Flash and Qwen3.8 27B.

	DeepSeek Flash			Qwen3.8 27B		
Source language	T	K	V	T	K	V
English	54.21	83.91	68.58	61.20	82.60	65.76
Czech	76.71	85.25	83.57	64.10	81.65	80.63
German	58.01	89.89	64.80	59.01	92.96	68.48
Dutch	73.19	85.29	78.24	71.82	80.92	80.94
French	68.31	87.55	65.85	63.60	89.95	58.89
Italian	54.60	54.23	53.98	60.46	68.06	65.50
Spanish	70.44	89.31	68.11	66.65	92.29	67.08
Portuguese	61.05	93.09	77.37	63.51	95.16	78.14
Turkish	64.43	82.75	59.61	61.99	92.00	58.03
Indonesian	73.50	92.14	80.25	65.32	85.66	71.57
Malay	62.97	97.91	80.72	63.22	90.97	78.34
Vietnamese	65.84	91.91	71.71	62.71	84.50	66.45
Russian	64.35	80.09	69.00	71.99	82.65	71.13
Ukrainian	58.88	95.77	63.14	62.27	92.14	62.14
Arabic	52.24	72.85	62.38	58.45	80.55	68.26
Hebrew	49.34	60.34	53.88	54.13	46.26	66.75
Hindi	65.89	75.21	69.40	60.37	83.16	67.27
Burmese	34.74	35.57	47.97	25.35	20.88	36.73
Thai	53.94	63.49	68.64	52.42	67.88	66.22
Japanese	59.58	69.00	52.96	64.44	73.32	65.25
Korean	59.34	61.79	59.71	63.30	71.43	68.96
Chinese	61.78	77.20	68.30	64.79	72.82	75.06

Alignment and variation across languages. Across the 12 models within each source language, Pearson correlations range from 0.9277 to 0.9917 for T–V and from 0.8635 to 0.9932 for T–K, showing strong alignment. GPT-6 Astra leads T on 16 of the 22 source languages, V on 15, and K on six. Within GPT-6 Astra, correlations across the 22 source-language means are 0.6820 for T–V and 0.7274 for T–K, lower than the cross-model correlations in Section 4.5.

Burmese has the lowest T for 11 of the 12 models, the lowest K for 11, and the lowest V for ten. For GPT-6 Astra, its Burmese scores are 51.22 in T, 51.29 in K, and 66.22 in V, compared with 78.92, 88.64, and 91.33 on English. The lower scores identify Burmese as a shared difficulty under the three rubric diagnostics; they do not isolate the source of the errors or exclude annotation and judge limitations.