OMNILINGUALGAIA2: EVALUATING THE MULTILINGUAL GAP IN FRONTIER AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic benchmarks aim to measure how well AI agents plan, reason, and act in realistic environments, yet they are almost exclusively English-only. As AI agents are increasingly deployed to linguistically diverse users, whether agentic competence measured in English generalises to other languages remains an open question. We introduce OMNILINGUALGAIA2, a machine-translated expansion of the GAIA2 agentic benchmark, with partial human-expert validation. OMNILINGUALGAIA2 covers ten target languages spanning five writing systems and includes a cross-lingually calibrated multilingual verifier. Evaluating seven proprietary and open-weight agents, we find a universal cross-lingual gap of 8.8–18.4 pp in pass@3. The gap varies substantially across agents, is concentrated in tool orchestration rather than quantitative reasoning, and persists with increasing model scale. A stratified error attribution estimates that 55% of the observed gap is due to agent-driven failures, which we further study through human-expert linguistic analysis. We identify morphological cue loss and amplified ambiguity as the primary failure mechanisms, especially in non-Latin-script languages. These findings demonstrate that strong English-language agentic performance does not reliably transfer across languages, and argue that multilingual agentic evaluation should become a standard component of reporting for globally deployed agents.

1 INTRODUCTION

AI agents—systems that plan, reason, and act autonomously within an environment to pursue a user's goals (Wooldridge & Jennings, 1995; Franklin & Graesser, 1997; Russell & Norvig, 2010)—have moved from objects of research curiosity to products deployed to millions of users worldwide, enabled by rapid advances in large language model (LLM) capabilities (Wang et al., 2024; Xi et al., 2025; Yang et al., 2025; Staufer et al., 2026). Those users do not all speak English, yet our ability to measure whether these agents actually work has barely kept pace with their deployment, especially beyond English. Even in English the ground is unsteady: developers who anticipated a 24% speedup from AI tools were instead slowed by 19% (Becker et al., 2025), and observed AI task coverage remains a fraction of theoretical estimates (Massenkoff & McCrory, 2026). Whether the competence we certify in English survives once an agent must act in another language is, at the time of writing, largely unmeasured. In English, this gap is being rapidly closed, with considerable work invested in agentic benchmarks (Mialon et al., 2024; Qin et al., 2024; Xie et al., 2024; Patil et al., 2025; Deng et al., 2026; Barres et al., 2025; Merrill et al., 2026). Beyond English, evaluation has barely followed: for frontier agentic systems, multilingual assessment remains largely confined to single-turn, knowledge-based question answering datasets such as GLOBAL-MMLU (Singh et al., 2025)[1], when it is performed at all[2].

To help close this gap, we present **OMNILINGUALGAIA2**, a high-quality machine-translated expansion of GAIA2 (Froger et al., 2026), paired with a multilingual verifier and a frontier multilingual agentic leaderboard. Our contributions are:

1. **A multilingual agentic benchmark.** We will release OMNILINGUALGAIA2, a machine-translated (MT) expansion of GAIA2 that covers four capabilities, namely adaptability,

[1] See for instance System Card: Claude Opus 4.7

[2] No multilingual performance reported in GPT-5.4 Thinking System Card

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.

ambiguity, execution, and search, across ten target languages spanning five writing systems while preserving the executable structure required by the verifier (Section 2). We select the translation system through a per-language quality evaluation and design an MT pipeline that enforces cross-surface terminology consistency across scenarios (Section 2.2).

2. **A human-calibrated multilingual verifier.** We localise the LLM-as-a-judge (LLMaaJ) component of the GAIA2 verifier and ensure its multilingual quality via calibration on a gold set of human-annotated English and machine-translated agent trajectories (Section 2.3).
3. **A cross-lingual multi-agent leaderboard and empirical characterisation of the gap.** We evaluate a heterogeneous cohort of seven AI agents spanning proprietary and open-weight models of various sizes and architectures, and find that: (i) the cross-lingual gap is universal in direction but agent-asymmetric in magnitude (8.8–18.4 pp); (ii) the capability ordering (search > execution > adaptability > ambiguity) is stable across languages—translation reshapes the level without reordering difficulty; (iii) agents shift their behavioural strategy in target languages, increasing exploratory reads while reducing state-changing writes; (iv) the cross-lingual gap is concentrated in tool-orchestration and final-response quality, not in quantitative or categorical reasoning; and (v) the gap does not close with model scale within a single family (Appendix E).
4. **An error taxonomy and gap attribution showing the gap is predominantly agent-driven.** We define a five-category verdict taxonomy. The resulting decomposition attributes 55% of the cross-lingual gap to genuine agent failures, 35% to translation defects, and 10% to verifier artefacts (Appendix G and section 5.1). Extending the analysis to the whole benchmark, we bound MT-induced contamination at only 6.4% of all scenario–language pairs.
5. **A human-expert linguistic analysis identifying failure mechanisms per error category.** Three native-speaker linguists independently inspect agent traces across the four capabilities and identify qualitative patterns across translation defects, verifier and measurement artefacts, and genuine agent failures. These include the consequential over-actions driven by morphological-cue loss—articles and plural marking—in `cmn`, `jpn`, and `ind` (Section 5.2).

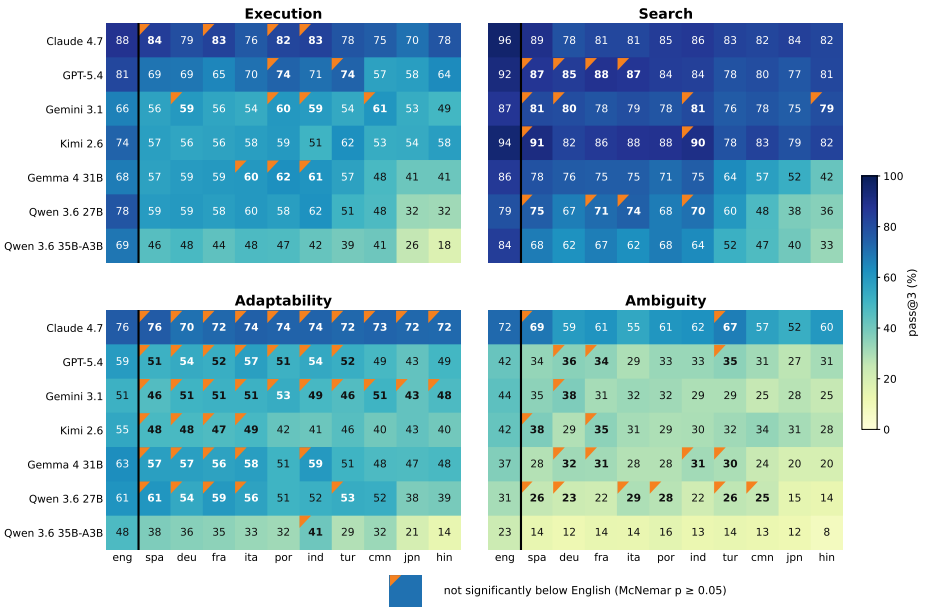

Figure 1: Per-capability × per-language pass@3 (%) across agents. Cross-lingually robust cells, i.e. target-language cells that are *not* significantly different from the corresponding English cell from the same (agent, capability) combination under a paired McNemar test ($p \geq 0.05$), are marked with an orange corner wedge.

2 OMNILINGUALGAIA2

2.1 SOURCE BENCHMARK

We build on GAIA2 (Froger et al., 2026), an agentic benchmark in which an agent operates within a simulated environment equipped with tools. We use its native *Mobile* environment, which comprises *apps* such as *Email* and *Calendar* and corresponding collections of *app states*, called *universes*. Each universe corresponds to a synthetic *persona* and captures a snapshot of their digital life through synthetic data such as emails and calendar events. A scenario specifies a task within this environment and the minimal set of write actions the agent must take, ordered logically and/or temporally, including the final answer. The verifier compares the actions the agent actually takes (*events*) with these *oracle events*. Events that contain free-form text, such as the final answer, undergo LLMaaJ verification to determine whether the scenario succeeds or fails. The initial version of OMNILINGUALGAIA2 covers four of GAIA2's five core capabilities: **adaptability**, **ambiguity**, **execution**, and **search**. We leave Time and the Noise and Agent2Agent augmentations for future work because extending coverage requires adapting and validating our multilingual construction and evaluation pipeline for temporal and asynchronous scenarios, injected environment perturbations, and app-agent communication.

2.2 TRANSLATING THE BENCHMARK

Translation pipeline. Translating a *verifiable* agentic task is more complex than translating a static question–answer pair: the translated environment must remain internally consistent for the verifier to admit exactly the intended solution. To construct OMNILINGUALGAIA2, we translate the language-dependent surfaces of every scenario into each target language: the task description and the natural-language portions of the universes and oracle events (*text spans*). We leave executable structure, such as identifiers and timestamps, untouched (*frozen spans*).

Figure 2 illustrates the three-stage pipeline. Since GAIA2 defines multiple scenarios over the same universe, we first aggregate the universes and translate their text spans. We then build a reference *term table* from these translated surfaces, primarily from the translated universes and augmented by a per-scenario extraction pass. The table identifies and consolidates entities that recur across surfaces, enforcing the consistency needed to avoid unsolvable tasks that refer to non-existent entities. Finally, we translate task descriptions and oracle-event text spans using the term table to constrain their rendering. A validation pass scans each assembled scenario and replaces any residual source-language span using the table, keeping the environment consistent with the verifier's expectations.

Translation system. We select GEMMA-4-31B-INSTRUCT (Gemma Team, 2026) as the translation system for all target languages because it is the highest-ranked open-weight model on the BOUQuET leaderboard (Andrews et al., 2025; Meta FAIR, 2026). We validate this choice by measuring translation quality through pairwise LLMaaJ comparisons of each candidate rendering with a reference translation, evaluated by two independent judges, each taking the majority verdict over three position-randomised trials. Per-language head-to-head comparisons with other shortlisted systems confirm this choice. We also test a reviewer stage with a different model, but find that it has little effect: it rewrites fewer than 9% of fields, the changes are almost entirely stylistic, and translation quality does not measurably improve, while per-scenario latency roughly doubles. See Appendix B for additional details.

2.3 MULTILINGUAL VERIFIER

GAIA2 evaluates every state-changing (*write*) action that the agent performs in the environment against oracle events via a multi-step *verifier*: it runs individual deterministic and natural-language *checkers*, with the latter invoking an LLMaaJ component that we call the *judge*. Extending the benchmark beyond English therefore requires localising not only the task descriptions, universe app states, and oracle events, but also this judge. Below, we describe its localisation and calibration for parity with English and correctness in the target languages; Appendix C provides additional details.

Localising the judge. Judge quality depends on two components: the **judge prompt** and the **judge model**. To evaluate them independently of agent performance, we use human-annotated scenario traces from GAIA2, each with a ground-truth pass/fail label. We translate the same traces with our

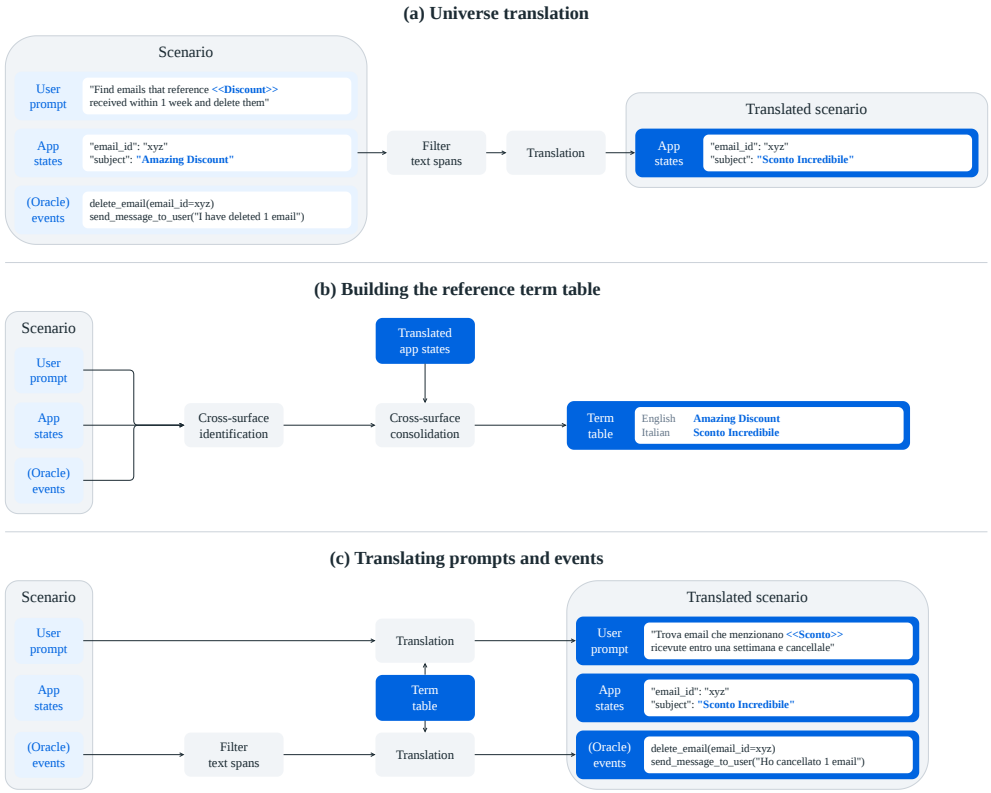

Figure 2: **The OMNILINGUALGAIA2 translation pipeline.** Each GAIA2 scenario is translated in three stages. **(a) Universe translation.** The LLM translator renders the language-dependent *text spans* of the universe app states into the target language, in calls batched by app. **(b) Building the term table.** A second LLM pass identifies entities that recur across surfaces and consolidates them into a *term table*, which records one rendering of each shared entity. **(c) Translating task descriptions and events.** The task descriptions and oracle events are translated under the constraints of the term table. A final pass substitutes any residual or divergent source-language span so that every surface refers to each entity identically—the cross-surface consistency the verifier requires to admit exactly the intended solution.

pipeline (Section 2.2), producing cross-lingual counterparts that let us evaluate the source and target languages on identical traces. On the prompt axis, we identify English-specific assumptions that make the original prompts unfit for translated content. For instance, the content checker is primed exclusively with English few-shot examples, so it can judge a faithful target-language write action as non-compliant for reasons unrelated to correctness. OMNILINGUALGAIA2 therefore includes localised judge prompts that remove this priming and generalise the natural-language sub-checks across target languages. On the model axis, we require a model that is open-weight and lightweight enough to score the entire multilingual leaderboard. GPT-OSS-120B (OpenAI, 2025) meets both requirements. As a mixture-of-experts model, it activates only 5.1B parameters per token, roughly an order of magnitude fewer than the dense LLAMA-3.3-70B-INSTRUCT (Meta AI, 2024) model used as the original paper's reference judge.

Calibrating the judge. We evaluate all four combinations of the two judge models and two prompt sets on the same human-annotated traces. We score each configuration against the human-majority label using Cohen's κ, which measures chance-corrected agreement. On English, no configuration reproduces the human verdicts measurably better than the others: all four achieve comparable agreement with the human labels (full-corpus κ within 0.71–0.73, rising to $\kappa \geq 0.93$ on the *LLM-adjudicated* slice). Crucially, the proposed configuration (GPT-OSS-120B, localised) agrees with the original paper's reference judge on almost all traces, with $\kappa = 0.981$. This agreement persists

under translation across all ten target languages. On the LLM-adjudicated slice, target-language Cohen's κ ranges from 0.886 (tur) to 0.971 (cmn), matching the English maximum of 0.971.

3 Experimental setup

Agents and models. We evaluate seven agents on OmniLingualGAIA2, each comprising an underlying model operated through a tool-use harness. The models span proprietary and open-weight systems and, among the latter, dense and mixture-of-experts architectures at different scales. The proprietary models are Claude-4.7-Opus (Anthropic, 2026), GPT-5.4 (OpenAI, 2026), and Gemini-3.1-Pro (Google DeepMind, 2026). The open-weight models are Gemma-4-31B-Instruct (Gemma Team, 2026), Kimi-2.6 (Kimi Team, 2026), and two Qwen3.6 variants: the mixture-of-experts Qwen-3.6-35B-A3B (Qwen Team, 2026c) and the dense Qwen-3.6-27B (Qwen Team, 2026b). To examine the effect of model scale, we also evaluate four sizes from the Qwen-3.5 family (Qwen Team, 2026a) (Appendix E).

Agent harness. All agents use the OpenClaw (OpenClaw Foundation, 2025) harness except Kimi-2.6, for which we use OpenCode (Anomaly, 2025), since we found that under OpenClaw, Kimi-2.6 scores near zero.[3]

Generation settings. We serve each open-weight model with a default 131K-token context window. When a rollout fails solely because its trajectory is truncated, we re-run it using the model's maximum context window and re-score it.[4] We enable thinking mode for the open-weight models but otherwise retain each model's default generation settings. For proprietary models, we use each provider's default reasoning-effort setting. To check whether this choice drives our results, we run a controlled, smaller-scale reasoning-effort ladder. The cross-lingual gap persists, largely undiminished, at the highest effort tested (Appendix H).

Metrics. Our main metric is pass@3: a scenario counts as solved if any of three independent rollouts passes the verifier. We also report pass@1, the mean per-rollout success rate used by GAIA2 (Froger et al., 2026), and $\text{pass}^{\text{all}}_3$, which requires all three rollouts to pass (Appendix D).

4 Results

4.1 Leaderboard

Table 1 reports pooled pass@3 scores across the four capabilities (adaptability, ambiguity, execution, search). Every agent's target-language average is lower than its English score, but the size of this gap varies substantially: from 8.8 pp in pass@3 for Gemini-3.1-Pro and 9.1 pp for Claude-4.7-Opus to 18.4 pp for Qwen-3.6-35B-A3B. The same ordering holds under pass@1, where Gemini-3.1-Pro again has the smallest pooled gap, at 7.7 pp. This robustness is relative: Gemini-3.1-Pro has the lowest English and multilingual performance of the three proprietary agents (62.2 pass@3 in English). The strongest open-weight model, Kimi-2.6, scores 4.3 pp above Gemini-3.1-Pro in English (66.5 versus 62.2), but only 0.8 pp above it on the target-language average (54.2 versus 53.4), resulting in a larger cross-lingual gap.

4.2 Results summary

The capability-level breakdown of Table 1 (Figure 1; exact numbers, 95% Wilson confidence intervals (CIs), and per-cell McNemar tests in Appendix D) shows that performance is strongest in English almost everywhere. The drop is concentrated in the three non-Latin-script languages (cmn, jpn, hin), partly because agents mishandle variation in entity forms when calling tools (Appendix J). The ordering of the capabilities—search > execution > adaptability > ambiguity—is stable across languages. The two strongest proprietary agents (Claude-4.7-Opus, GPT-5.4) lose the least on

[3] We traced this to the harness dropping its native streaming `reasoning_content` and `tool_calls`.

[4] Truncation is concentrated in three non-Latin-script languages, affecting up to 20% of cmn, 34% of jpn, and 44% of hin rollouts.

Agent	Exec.	Search	Adapt.	Ambig.	Avg.
			English		
CLAUDE-4.7-OPUS	**88.1**	**95.6**	**75.6**	**71.6**	**82.7**
GPT-5.4	81.2	92.4	58.8	41.4	68.4
GEMINI-3.1-PRO	66.2	86.9	51.2	44.4	62.2
KIMI-2.6	73.8	94.3	55.0	42.8	66.5
GEMMA-4-31B-INSTRUCT	67.5	85.6	63.1	36.9	63.3
QWEN-3.6-35B-A3B	68.8	86.2	48.1	23.1	56.6
QWEN-3.6-27B	77.5	79.4	61.2	30.6	62.2
			Target-language average		
CLAUDE-4.7-OPUS	**78.8**	83.0	**73.0**	**59.7**	**73.6**
GPT-5.4	67.1	83.5	51.4	31.8	58.5
GEMINI-3.1-PRO	56.1	78.3	48.9	30.4	53.4
KIMI-2.6	56.5	**84.8**	44.5	31.1	54.2
GEMMA-4-31B-INSTRUCT	55.2	69.5	53.4	27.5	51.4
QWEN-3.6-35B-A3B	41.8	65.6	31.8	13.6	38.2
QWEN-3.6-27B	55.2	69.4	51.9	24.4	50.2

Table 1: Per-capability pass@3 (%) across agents; best per column in **bold**. Exec., Adapt., Ambig., and Avg. denote execution, adaptability, ambiguity, and average, respectively. The target-language average pools the ten target languages.

adaptability; every agent loses most on *ambiguity*, the hardest capability in every language. Additional rollouts rarely recover the cross-lingual loss. A Gemma–Qwen agent–translator ablation also finds no detectable family-match *family-match* advantage arising from the use of GEMMA-4-31B-INSTRUCT in the translation pipeline (Appendix B.3). Appendix F.1 provides the full analysis.

For two proprietary agents (CLAUDE-4.7-OPUS, GPT-5.4) and the strongest open-weight agent (KIMI-2.6), per-scenario step counts and tool choices show how behaviour shifts under translation. On scenarios where performance regresses, CLAUDE-4.7-OPUS makes markedly *fewer* tool calls than the other agents, a "give up early" signature. By contrast, GPT-5.4 systematically does more work in the target languages, but that additional work does not yield accuracy gains. All three agents change *which* tools they use far more than the pooled average suggests. The direction of the change is consistent: in the target languages, a larger share of the tool budget shifts from *state-changing writes* to *exploratory reads*. Appendix F.2 provides the full quantitative breakdown.

We label each failed rollout by the first verifier check it fails and report *absolute* failure rates. The cross-lingual regression does *not* arise from quantitative or categorical reasoning: for all three agents, every such delta falls within ± 0.35 pp. Instead, the regression is concentrated in tool-count mismatch and textual-reply rejection, which together account for essentially all of the rise in target-language failure. The cross-lingual gap is therefore a tool-orchestration and final-response-quality effect, not a numeric- or categorical-reasoning one. Appendix F.3 provides the full per-check rates and methodology.

5 ERROR ATTRIBUTION: AUTOMATIC AND LINGUISTIC ANALYSIS

With linguists and language experts, we first define five error categories: infrastructure failures (crashes or missing judgements), translation defects, verifier artefacts, agent failures, and inconclusive cases (Appendix G). We then apply this taxonomy in two complementary studies: large-scale diagnostic-agent triage and targeted human linguistic analysis.

5.1 AUTOMATIC ESTIMATION

An LLM-agent protocol assigns each failure to a taxonomy category from the per-rollout artefacts persisted by the harness and benchmark orchestrator. It compares a failing target-language rollout with a successful English rollout of the same scenario: the oracle events define success, while the English rollout shows how the same agent achieved it. Rollouts without a gradable attempt (8.5%) are excluded throughout; Appendix I describes the protocol in full.

Gap-representative stratification. The severity of a regression is itself informative. A deterministic translation defect breaks *every* target rollout; an agent slip or flaky judge breaks only *some*. Triaging complete collapses alone (English 3/3, target 0/3) would therefore over-represent translation defects. We instead stratify regressions by within-scenario determinism, then triage a language- and capability-balanced sample from each stratum at the level of individual failing rollouts. Finally, we reweight the strata by their true failing-rollout volumes and compute scenario-clustered bootstrap intervals.

The two strata differ sharply (Table 17). Deterministic regressions are translation-driven (59%), whereas stochastic regressions—which account for the larger share of the gap—are agent-driven (75%). After reweighting (Table 18), 55.4% (95% CI 50.3–60.7) of the gap is attributable to genuine agent failure, 34.5% (29.5–39.1) to translation defects, and 10.1% (7.1–13.5) to verifier artefacts. A complete-collapse sample would instead yield a translation-dominated $\approx 55\%/33\%$ split. The agent-driven share is largest for Turkish and Japanese and smallest for Portuguese and Spanish (Appendix I).

From regressions to the whole benchmark. Conditioning on English *passing* leaves the triage blind to scenarios that fail in both languages. This is precisely where a defect could depress a target score without producing a visible regression. For the cross-lingual scores, the most consequential MT failure mode is a translation defect that makes the target version *unsolvable*. The 74.5% of all 6,056 English×target scenario–language pairs solved at least once in the target language therefore provide direct evidence against this severe failure mode (Table 19).

For pairs the target never solves, we group by English outcome: pass→fail, partial→fail, and fail→fail. Their translation-defect rates fall from 59% to 16% to 10% (95% CI 5.6–16.9 for fail→fail), respectively. Defects thus concentrate where they are already *visible* as clean regressions. Moreover, a defect in the fail→fail cell cannot bias the cross-lingual ranking, since a scenario that also fails in English contributes nothing to the gap. Weighted by cell size, translation renders only 6.4% of all pairs unsolvable in the target. These defects cluster around a handful of oracle- and keyword-mistranslation bugs, leaving a small, localised and fixable contamination floor. The gap is therefore robust to MT-induced unsolvability and predominantly reflects genuine agent capability.

Human validation. Five human reviewers re-adjudicated a stratified sample of triaged failures through a purpose-built review interface. They judged the fault side using only decision-relevant evidence: the source and translated task descriptions, oracle event, judge rationale, and agent answer. They reviewed **149** items in seven languages, of which 9 were undecidable; among the remaining **140**, they confirmed the automatic label in **128**, corresponding to 91.4% agreement (95% CI 85.6–95.0). Agreement was comparable across all three sides, and rejections crossed category boundaries in no systematic direction (Appendix K).

5.2 Linguistic analysis

Three linguists independently inspected agent traces, translated task descriptions, and universe content for 16 scenarios spanning all four capabilities. They were native speakers of Indonesian, Mandarin Chinese, and Spanish, respectively, and all were proficient in English; the Mandarin speaker was also proficient in Japanese. We organise their observations around the three non-infrastructure fault sides of the taxonomy (Appendix G), mirroring Section 5.1. Table 15 presents six representative cases.

Translation defects (T). In several task descriptions, translation neutralises the word-order or morphological cues that disambiguate the English, forcing a single and potentially wrong reading. For example, "whose titles" was rendered as *job title*, rather than *event title*, in every target language (Case 1, T3). Other translations appear faithful while silently dropping constraints. A directional preposition in "booked a cab *to*" was dropped or weakened in every target language, removing the constraint that determines the correct answer (Case 2, T3). The most consequential failures arise from the loss of morphological cues: in article-less, number-neutral languages (`cmn`/`jpn`/`ind`), the singular "my contact" is interpreted as a generic set, leading the agent to delete *both* matching contacts instead of asking for clarification (Case 6, T3).

Verifier artefacts (K). Judge strictness varies by language, behaves stochastically, and can hallucinate. The judge rejects correct target-language content while accepting identical English content;

in one case, it cited agent text that was never produced (Case 4, K1). The deterministic write-action-count checker creates a separate problem by penalising harmless self-corrections: it rejects any mismatch with the oracle's write-action count even when the final state is correct, as when an agent omits an attendee, deletes the event, and recreates it correctly (Cases 3–5). Although the taxonomy records these as extra-action agent errors (A3), their correct end state makes them largely *measurement* artefacts of end-state-blind grading.

Agent failures (A). Genuine agent errors are nevertheless present and, in aggregate, dominant. Consistent with Section 5.1, the sample includes clean slips attributable to neither MT nor the verifier: a Spanish rollout skips a required save step despite naming it in its own reasoning (Case 3, A2), while a retrieval miss returns only one of two matching contacts (Case 6, `spa`). Extra writes, however, are not uniformly benign. An over-action that changes the final state is a true A3 regression, unlike the idempotent self-corrections penalised above. The two must be separated before an action-count failure is interpreted as a capability gap.

Overall, the qualitative sample matches the automatic decomposition. Translation defects and verifier artefacts are real but bounded, while translation-independent agent errors surface even at this small scale.

6 Related work

Multilingual agentic evaluation. Efforts to move agentic evaluation beyond English have so far been narrow along at least one of three axes: the languages covered, the range of agents evaluated, or the fidelity of the verifier under translation. Some benchmarks isolate a single sub-capability or set of locales, such as web navigation (Wang et al., 2025), function calling for regionalised customer service (Sales Almeida et al., 2025) or voice assistants (Kulkarni et al., 2025), bilingual telecom troubleshooting (Bariah et al., 2026), or cross-lingual tool-calling robustness (Luo et al., 2026). Others evaluate a broader multilingual agent suite but retain a narrower analytical focus (Hofman et al., 2026).

Three recent efforts are closest to our setting. SEATauBench (Nguyen et al., 2026) adapts τ^2-Bench into five Southeast Asian languages, while PolyWorkBench (Li et al., 2026) constructs a native multilingual long-horizon benchmark from scratch. The most direct comparison is GAIA-v2-LILT (Kim et al., 2026), which localises GAIA into five target languages through a translate-then-adapt pipeline that modifies task semantics to fit each locale.

OmniLingualGAIA2 takes a different route: it expands an existing benchmark through MT. Unlike synthetic-generation approaches (Levi & Kadar, 2025; Shi et al., 2026; Xie et al., 2026), an MT-based expansion can scale linguistically while preserving the executable structure of the original verifier. Our pipeline follows the hybrid MT-then-expert-review approach of MMLU-ProX (Xuan et al., 2025). Its design is also informed by the Agentic Benchmark Checklist (Zhu et al., 2025), which documents systematic failures of outcome and task validity in widely used agentic benchmarks. We address these risks by calibrating the verifier against human labels (Section 2.3) and decomposing the observed gap into translation defects, verifier artefacts, and agent failures, including a bounded contamination floor (Section 5.1).

Against this landscape, OmniLingualGAIA2 is, to our knowledge, the first extension of GAIA2 that **(i)** covers ten target languages across five writing systems (Appendix A) while preserving the executable verifier end-to-end; **(ii)** reports a paired comparison across a heterogeneous cohort of proprietary and open-weight agents using a shared, calibrated cross-lingual judge; and **(iii)** decomposes the observed gap into translation-, verifier-, and agent-driven factors, extending the analysis from isolated tool calls to full agent trajectories and complementing it with human-expert linguistic analysis.

Multilingual LLM-as-a-judge. OmniLingualGAIA2, like GAIA2, uses a model-based verifier to grade agents' natural-language write actions. The verifier's cross-lingual reliability is therefore a central concern. LLM judging parallels MT quality estimation, both prompted (Kocmi & Federmann, 2023) and learned (Rei et al., 2022; Juraska et al., 2025), but even monolingual judges are subject to position, verbosity, and self-preference biases (Zheng et al., 2023; Koo et al., 2024; Panickssery et al., 2024).

These problems can worsen under translation. Judge–human agreement varies across languages, degrading for lower-resource ones (Hada et al., 2024; Son et al., 2024; Gureja et al., 2025); order consistency nears chance for some languages (KC, 2026); and judges exhibit a *translationese* bias, favouring back-translated over human-authored responses (Zhang et al., 2026). We thus recalibrate the GAIA2 LLMaaJ scaffold against human-annotated multilingual trajectories (Section 2.3). We measure inter-rater agreement with standard statistics (Cohen, 1960; Artstein & Poesio, 2008; Landis & Koch, 1977; Krippendorff, 2019) and report language-conditional confidence intervals.

7 Conclusion

We present OmnilingualGAIA2, a machine-translated multilingual expansion of the GAIA2 agentic benchmark spanning ten priority target languages. We pair it with a localised and human-calibrated verifier and evaluate a heterogeneous seven-agent cohort spanning proprietary and open-weight models with both dense and mixture-of-experts architectures. Our main findings are as follows. First, the cross-lingual gap is universal in direction but agent-specific in magnitude (8.8–18.4 pp in pass@3), and no agent is uniformly robust: the proprietary systems narrow the gap without closing it. The two success metrics distinguish the systems further: Gemini-3.1-Pro is the most self-consistent (smallest $\text{pass}^{\text{all}}_3$ gap), while Claude-4.7-Opus retains high any-of-three accuracy in target languages but loses per-rollout reliability. Second, the gap is predominantly agent-driven: a stratified error attribution assigns 55% of it to genuine agent failures, 35% to translation defects, and 10% to verifier artefacts. A benchmark-wide bound leaves only 6.4% of scenario–language pairs MT-unsolvable, and the gap concentrates on tool orchestration and response quality rather than quantitative or categorical reasoning. Third, this gap does not close with model scale: a same-family size ladder shows the English–target-language delta widening from $+8$ to $+13$ pp as the number of active parameters grows. Fourth, agents adopt a more hesitant execution strategy in target languages, foraging longer over the environment and committing later and less decisively (the write-action share of tool calls falls by 1–2.5 pp); this shift is the behavioural signature of degraded comprehension: it accompanies the mis-counted write actions behind the execution-side failures rather than reflecting a successful adaptation. Finally, linguistic analysis identifies morphological cue loss and amplified ambiguity as primary failure mechanisms, particularly in non-Latin-script languages.

Taken together, our results argue that multilingual agentic evaluation belongs alongside the evaluation of multilingual understanding as a standard part of reporting for globally deployed agents. Two priorities follow. The first is to broaden coverage along both axes: to the three capabilities deferred in this work and to more target languages, ideally focusing on the lower-resource end, where we posit that the cross-lingual gap is widest. The second is to harden the translation pipeline (Section 2.2) against the defect mechanisms surfaced by our error analysis (Section 5.2), driving down the benchmark's 6.4% translation-induced unsolvability floor.

Limitations

- **Excluded capabilities.** We cover four of GAIA2's five core capabilities, leaving Time and the Noise and Agent2Agent augmentations for future work pending multilingual validation of temporal and asynchronous scenarios, injected perturbations, and app-agent communication.
- **Translation quality and audit coverage.** OmnilingualGAIA2 is automatically translated without full-corpus human post-editing, and its quality signals do not replace professional per-language review. Residual defects concentrate in non-Latin, morphologically rich languages, while our human audit covers only selected languages and regression samples (Section 5); affected scores should be interpreted accordingly.
- **Judge reliability ceiling.** Moderate agreement among the human annotators bounds our calibration: high judge agreement shows parity with an imperfect reference, not absolute correctness. Raising this ceiling requires higher-agreement re-annotation.
- **System configuration.** We use OpenClaw for six agents and OpenCode for Kimi-2.6; without a harness ablation, another harness—including one co-developed with the model—could reorder the leaderboard. The proprietary systems use provider-default reasoning effort. Our controlled ladder suggests that reasoning budget does not drive the multilingual gap, but covers only one model and three languages.

AI USE STATEMENT

Generative AI tools were used in this work as an object of study and as a component of the research pipeline: the target-language versions of the benchmark were produced by machine translation and the multilingual verifier is LLM-based, as described in Section 2.2 and Section 2.3; both were validated against human-expert judgements (Appendix C). An LLM-agent protocol also supported the error-attribution analysis described in Section 5.1. Generative AI also assisted with implementing the translation pipeline and benchmark execution code. For the manuscript, it was used only to double-check claims and to polish the writing. All AI-assisted output was reviewed by the authors, and we take responsibility for the final content of this work, including all text, claims, and artefacts.

ETHICS STATEMENT

OMNILINGUALGAIA2 is an evaluation benchmark built on GAIA2 and contains no personal data beyond the synthetic content in the original scenarios. Our human analyses involved three expert linguists (Section 5.2) and five reviewers who validated the automatic error attribution (Section 5.1). All were members of, or affiliated with, the authors' research team, contributed as part of their regular duties, and examined only synthetic benchmark content and agent outputs. The verifier calibration reuses existing human annotations of GAIA2 traces; no new judgements were collected. Our goal is to make cross-lingual performance gaps measurable: documenting where agents underperform for non-English users is a prerequisite for serving those users equitably. Machine-translated benchmarks can nevertheless introduce translationese and cultural mismatches. We therefore do not interpret OMNILINGUALGAIA2 scores as a measure of culturally appropriate behaviour, which we leave to future work.

REPRODUCIBILITY STATEMENT

All details needed to reproduce our results are given in the paper and its appendices: the machine-translation and quality-control pipeline in Section 2.2 and Appendix B, the language inventory in Appendix A, and the evaluation setup (models, decoding settings, and pass@k protocol) in Section 3. The exception is the verifier calibration (Section 2.3): we report its full protocol and agreement statistics in Appendix C, but not the underlying human labels. Complete per-language results are reported in Appendix D. The benchmark and the evaluation code will be released publicly.

REFERENCES

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27515–27535, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1400. URL https://aclanthology.org/2025.emnlp-main.1400/.

Anomaly. OpenCode: an open-source coding agent. https://opencode.ai, 2025. GitHub repository: https://github.com/anomalyco/opencode.

Anthropic. Claude Opus 4.7 system card. Technical report, Anthropic, 2026. URL https://www.anthropic.com/claude-opus-4-7-system-card.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. URL https://aclanthology.org/J08-4004.

Lucas Bandarkar, Alan Ansell, and Trevor Cohn. Large reasoning models struggle to transfer parametric knowledge across scripts, 2026. URL https://arxiv.org/abs/2603.17070.

Lina Bariah, Brahim Mefgouda, Farbod Tavakkoli, Enrique Molero, Louis Powell, and Merouane Debbah. TelcoAgent-Bench: A multilingual benchmark for telecom AI agents, 2026. URL https://arxiv.org/abs/2604.06209.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity, 2025. URL https://arxiv.org/abs/2507.09089.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa R Kundurthy, Sean M. Hendryx, Zifan Wang, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. SWE-Bench Pro: Can AI agents solve long-horizon software engineering tasks? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uEVTdoAbnK.

Stan Franklin and Art Graesser. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Intelligent Agents III: Agent Theories, Architectures, and Languages*, volume 1193 of *Lecture Notes in Computer Science*, pp. 21–35. Springer, 1997. doi: 10.1007/BFb0013570. URL https://doi.org/10.1007/BFb0013570.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9gw03JpKK4.

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

Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-RewardBench: Evaluating reward models in multilingual settings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43–58, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.3. URL https://aclanthology.org/2025.acl-long.3/.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. METAL: Towards multilingual meta-evaluation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2280–2298, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.148. URL https://aclanthology.org/2024.findings-naacl.148/.

Omer Hofman, Jonathan Brokman, Oren Rachmil, Shamik Bose, Vikas Pahuja, Toshiya Shimizu, Trisha Starostina, Kelly Marchisio, Seraphina Goldfarb-Tarrant, and Roman Vainshtein. MAPS: A multilingual benchmark for agent performance and security. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 821–845, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.42. URL https://aclanthology.org/2026.findings-eacl.42/.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 957–968, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.70. URL https://aclanthology.org/2025.wmt-1.70/.

Shreyas KC. BabelJudge: Measuring LLM-as-a-judge reliability across languages and agent trajectories, 2026. URL https://arxiv.org/abs/2606.22329.

Yunsu Kim, Kaden Uhlig, and Joern Wuebker. GAIA-v2-LILT: Multilingual adaptation of agent benchmark beyond translation. In *Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026)*, pp. 140–148, San Diego, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.mellm-1.13. URL https://aclanthology.org/2026.mellm-1.13/.

Kimi Team. Kimi K2.6: Advancing open-source coding. Moonshot AI blog, 2026. URL https://www.kimi.com/blog/kimi-k2-6.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19/.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, 2024. URL https://aclanthology.org/2024.findings-acl.29/.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 4 edition, 2019.

Mayank Kulkarni, Vittorio Mazzia, Judith Gaspers, Chris Hench, and Jack FitzGerald. MASSIVE-Agents: A benchmark for multilingual function-calling in 52 languages. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20193–20215, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1099. URL https://aclanthology.org/2025.findings-emnlp.1099/.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Elad Levi and Ilan Kadar. IntellAgent: A multi-agent framework for evaluating conversational AI systems, 2025. URL https://arxiv.org/abs/2501.11067.

Hongliang Li, Yijin Liu, Zhiwei Zhang, Zihe Liu, Xinyue Lou, Jinan Xu, Fandong Meng, and Kaiyu Huang. PolyWorkBench: Benchmarking LLM agents for cross-lingual long-horizon workflows, 2026. URL https://arxiv.org/abs/2607.06008.

Zheng Luo, T Pranav Kutralingam, Ogochukwu N Okoani, Wanpeng Xu, Hua Wei, and Xiyang Hu. Lost in execution: On the multilingual robustness of tool calling in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44059–44077. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2039. URL https://aclanthology.org/2026.acl-long.2039/.

Maxim Massenkoff and Peter McCrory. Labor market impacts of AI: A new measure and early evidence, 2026. URL https://www.anthropic.com/research/labor-market-impacts.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Mike A Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a7Qa4CcHak.

Meta AI. Llama 3.3 70B Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md, 2024. Released 2024-12-06.

Meta FAIR. BOUQuET leaderboard. https://huggingface.co/spaces/facebook/bouquet, 2026. Accessed September 2026.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2311.12983.

My Chiffon Nguyen, Aulia Adila, Saksorn Ruangtanusak, Kittiphat Leesombatwathana, Vissuta Gunawan Lim, Patomporn Payoungkhamdee, and Samuel Cahyawijaya. SEATauBench: Progressively adapting tool-agent-user evaluation into low-resource Southeast Asian languages, 2026. URL https://arxiv.org/abs/2606.28715. Findings of EMNLP 2026.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenAI. GPT-5.4 Thinking system card. Technical report, OpenAI, 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking.

OpenClaw Foundation. OpenClaw: an open-source, local-first AI assistant harness. https://github.com/openclaw/openclaw, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.60. URL https://aclanthology.org/2022.wmt-1.60/.

Stuart J Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, Inc., 3rd edition, 2010.

Thales Sales Almeida, João Guilherme Alves Santos, Thiago Laitz, and Giovana Kerche Bonás. Ticket-Bench: A kickoff for multilingual and regionalized agent evaluation, 2025. URL https://arxiv.org/abs/2509.14477.

Dingfeng Shi, Jingyi Cao, Qianben Chen, Weichen Sun, Weizhen Li, Hongxuan Lu, Fangchen Dong, Tianrui Qin, King Zhu, Minghao Liu, Yuchen Eleanor Jiang, Jian Yang, Ge Zhang, Jiaheng Liu, Changwang Zhang, Jun Wang, and Wangchunshu Zhou. TaskCraft: Automated generation of agentic tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UJFCyrYM1V.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. MM-Eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models, 2024. URL https://arxiv.org/abs/2410.17578.

Leon Staufer, Kevin Feng, Kevin Wei, Luke Bailey, Yawen Duan, Mick Yang, A Pinar Ozisik, Stephen Casper, and Noam Kolt. The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, pp. 1536–1576. ACM, 2026. doi: 10.1145/3805689.3806728. URL https://arxiv.org/abs/2602.17753.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Peng Wang, Ruihan Tao, Qiguang Chen, Mengkang Hu, and Libo Qin. X-WebAgentBench: A multilingual interactive web benchmark for evaluating global agentic system. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.988/.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Jingxu Xie, Dylan Xu, Xuandong Zhao, and Dawn Song. AgentSynth: Scalable task generation for generalist computer-use agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CoBxmXThM6.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.79. URL https://aclanthology.org/2025.emnlp-main.79/.

Jeremy Yang, Noah Yonack, Kate Zyskowski, Denis Yarats, Johnny Ho, and Jerry Ma. The adoption and usage of AI agents: Early evidence from Perplexity, 2025. URL https://arxiv.org/abs/2512.07828.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. Mitigating translationese bias in multilingual LLM-as-a-judge via disentangled information bottleneck. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.10351.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy K Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet S Sekhon, Jacob Steinhardt, Sarah Schwettmann, Arvind Narayanan, Matei Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices in building rigorous agentic benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=E58HNCqoaA.

A LANGUAGES

Table 2 specifies the language coverage of related multilingual agentic datasets, including ours.

Language	MAPS	GAIA-v2-LILT	OMNILINGUALGAIA2
Arabic	✓	✓	–
German	✓	✓	✓
Hindi	✓	✓	✓
Korean	✓	✓	–
Portuguese	✓	✓	✓
Italian	✓	–	✓
Japanese	✓	–	✓
Spanish	✓	–	✓
Russian	✓	–	–
Hebrew	✓	–	–
Mandarin Chinese	✓	–	✓
French	–	–	✓
Indonesian	–	–	✓
Turkish	–	–	✓

Table 2: Language coverage comparison between MAPS (Hofman et al., 2026), GAIA-v2-LILT (Kim et al., 2026), and our benchmark.

B THE TRANSLATION PIPELINE

In this section we report the results of the empirical studies we carry out that motivate the design of the final translation pipeline we adopt to produce the benchmark, as described in Section 2.2. In particular, these studies motivate the choice of GEMMA-4-31B-INSTRUCT as the translator and the choice to run it in a single translator-only pass with no downstream reviewer. Both decisions rest on the same pairwise protocol. Every translatable span produced by a candidate configuration is compared against the corresponding span of a *reference translation*, which is treated as the gold rendering of the same content. The comparison is evaluated by an LLM judge that returns one of three outcomes: *equivalent*, *candidate-wins*, or *reference-wins*. In particular, we consider a fixed sample of 200 spans, stratified by app, and score each (translator, language) combination against two LLM judges, GPT-5.4 and GEMINI-3.1-PRO; each judge's verdict is the majority over three position-randomised trials (thus controlling for the known first-position bias of pairwise LLM judges). A one–one–one split has no majority and is left unassigned, so the two judges together contribute 400 nominal comparisons per combination but occasionally fewer assigned verdicts. We report *candidate-win* rate (judge prefers the candidate over the reference, higher is better), *reference-win* rate (lower is better), and the equivalent share.

Both candidate translators run under a 16K-token context budget, within which each per-scenario prompt must accommodate the source spans, the generated translation, and the term table of Section 2.2. Since the raw term table—up to roughly a thousand entries per universe—would overflow this budget on its own, we cap it at its 200 shortest-source entries; a reproducibility check on spa finds that the capped and uncapped tables yield GEMMA-4-31B-INSTRUCT translations differing by under 2.5 pp on every judged cell, confirming the cap is quality-neutral. Within the prompt this table is rendered inline as the `Terminology` block shown in Figure 3, and is omitted for scenarios that share no terms across surfaces. Both translators are executed with the reasoning ("thinking") mode disabled: left enabled, it consistently exhausts the 16K-token budget on internal reasoning traces before emitting any parseable translation.

B.1 TRANSLATOR CHOICE

We evaluate the two top-ranking open-weight systems on BOUQuET for our target languages: GEMMA-4-31B-INSTRUCT and QWEN-3.6-27B, scoring translation across all ten languages from both systems against the reference translations under the protocol above. Table 3 reports the per-language verdict shares, and Figure 4 visualises them.

We find that GEMMA-4-31B-INSTRUCT dominates QWEN-3.6-27B on every one of the ten target languages: its reference-win rate is lower and its candidate-win rate higher on all ten. The margin is

```
Translate the following text from {src_lang} to {tgt_lang}.

### Terminology (use these exact translations for the following terms):
- "{source span 1}" -> "{target rendering 1}"
- "{source span 2}" -> "{target rendering 2}"
  ...

### Text to translate:
{src_text}

Keep any proper nouns, person names, IDs, email addresses, URLs, file names,
and technical identifiers unchanged.
When the text contains quoted references to app content
(conversation titles, email subjects, event titles),
use the exact translations from the terminology section above if provided.

Respond with ONLY the translated text, no explanations or comments.
```

Figure 3: The prompt supplied to the translator for every text span. `{src_lang}` and `{tgt_lang}` are the source and target language and `{src_text}` is the span to translate; the `Terminology` block is the per-scenario term table (Section 2.2) rendered inline (placeholder entries shown; the pipeline joins each pair with a "→" arrow), present only when the scenario shares terms across surfaces. Oracle tool-call arguments are translated with a structurally identical variant that additionally conditions on the task description and returns a JSON object keyed by argument index.

	GEMMA-4-31B-INSTRUCT			QWEN-3.6-27B		
Lang.	cand	ref	equiv	cand	ref	equiv
spa	**28.5**	**18.2**	53.2	29.5	30.5	39.8
por	**19.8**	**19.8**	60.0	14.2	37.5	47.8
fra	**21.5**	**24.5**	53.8	13.5	39.5	46.8
ita	**26.8**	**25.5**	47.5	19.2	41.0	39.2
deu	**24.5**	**20.5**	54.0	15.0	35.5	48.2
hin	**33.8**	**20.5**	45.2	19.2	49.8	30.5
cmn	**24.0**	**28.2**	45.8	20.8	30.8	46.2
jpn	**27.2**	**23.0**	49.2	19.2	42.8	37.8
tur	**32.2**	**16.8**	50.8	12.8	48.5	37.2
ind	**28.2**	**21.8**	50.0	15.8	46.2	37.5

Table 3: Translator head-to-head: pooled verdict shares (%, $n = 400$ nominal comparisons per cell) for GEMMA-4-31B-INSTRUCT and QWEN-3.6-27B against the reference translations, per target language. Lang., cand, ref, and equiv denote language, candidate wins, reference wins, and equivalent verdicts, respectively; rare no-majority comparisons remain in the denominator, so the three displayed shares can sum to slightly less than 100%. GEMMA-4-31B-INSTRUCT wins on every language; **bold** marks the winner's candidate-win (higher better) and reference-win (lower better) columns.

narrowest on the closest-to-saturation Latin-script targets (a -6 pp reference-win gap on `spa`) and widens sharply on the non-Latin and morphologically heavy targets (-32 pp on `tur`, -29 pp on `hin`), where surface-fidelity errors are harder to recover from. GEMMA-4-31B-INSTRUCT consistently wins on all ten languages and therefore we adopt it as our main and only translation system.

B.2 REVIEWER STAGE

We further investigate whether adding a reviewer stage, in which a second model flags and post-edits the translator's output, improves OMNILINGUALGAIA2; we find that it does not.

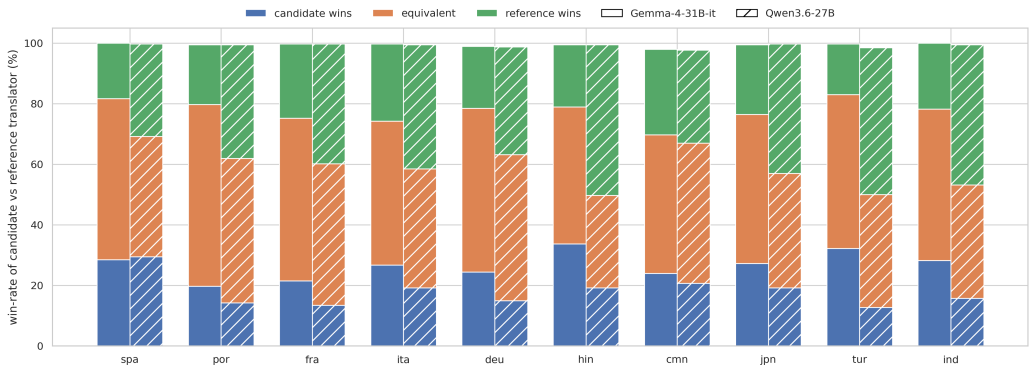

Figure 4: Translator head-to-head. Pooled breakdown of 400 nominal comparisons for GEMMA-4-31B-INSTRUCT (solid bars) and QWEN-3.6-27B (hatched bars) on each of the ten target languages, each translation judged pairwise against the reference translation. The reference-wins share (top segment) is smaller for GEMMA-4-31B-INSTRUCT on every language, and the gap widens on the non-Latin and morphologically heavy targets. Bars can end slightly below 100% because no-majority comparisons are left unassigned.

Fixing the translator to GEMMA-4-31B-INSTRUCT, we consider a reviewer stage with three different configurations: GPT-OSS-120B as a cross-family reviewer on all ten languages, GEMMA-4-31B-INSTRUCT itself as a self-review baseline on a three-language subset (`spa`, `cmn`, `ind`), and QWEN-3.6-27B as an alternate cross-family reviewer on the same subset. For each configuration we measure the change in candidate-win rate against the reviewer-free GEMMA-4-31B-INSTRUCT-only baseline, $\Delta = \hat{p}_{\text{reviewed}} - \hat{p}_{\text{baseline}}$.

Every one of the sixteen reviewer configurations is statistically indistinguishable from the reviewer-free baseline, since all the 95% confidence intervals on Δ cross zero (see Table 4). On the ten-language GPT-OSS-120B sweep the smallest uncorrected p-value is 0.47; pooling all ten cells gives a mean $\Delta = +1.0$ pp with a 95% CI of $[-1.0, +3.0]$ pp. We evaluate two other reviewer models on a reduced sample of languages and find that the choice of reviewer does not significantly alter the outcome.

Manual inspection of the edits confirms the statistical picture: the reviewer rewrites under 9% of fields, with changes that are almost entirely stylistic (gender inclusion, casing, idiom normalisation) rather than error repairs. We therefore commit to a translation pipeline for OMNILINGUALGAIA2 with a single translator-only pass, eliminating a second model pass that roughly doubles per-scenario latency for no measurable quality gain.

B.3 TRANSLATOR-FAMILY CONFOUND

Because GEMMA-4-31B-INSTRUCT serves both as the model in the translation pipeline (Section 2.2) and as one of the evaluated agents—and QWEN-3.6-27B is likewise evaluated—a natural concern is a *family-match* confound: an agent might enjoy a home-field advantage on data translated by its own model family, inflating its apparent capability. We test this directly with a 2×2 ablation crossing agent $\in$ {GEMMA-4-31B-INSTRUCT, QWEN-3.6-27B} with translator $\in$ {GEMMA-4-31B-INSTRUCT-MT, QWEN-3.6-27B-MT}, where QWEN-3.6-27B-MT retranslates the identical scenarios through the same pipeline. The design spans all ten target languages at 1,920 rollouts per cell (40 cells, 262K-token context), and we form the difference-in-differences (DiD)

$$\begin{aligned}\text{DiD} = &\left[G_{\text{GEMMA-4-31B-INSTRUCT-MT}} - G_{\text{QWEN-3.6-27B-MT}}\right] \\ &- \left[Q_{\text{GEMMA-4-31B-INSTRUCT-MT}} - Q_{\text{QWEN-3.6-27B-MT}}\right],\end{aligned} \tag{1}$$

where G and Q are the GEMMA-4-31B-INSTRUCT and QWEN-3.6-27B agent scores; DiD > 0 would indicate a GEMMA-4-31B-INSTRUCT home-field advantage.

The agent gap is essentially invariant to which family translated the evaluation data (Table 5): the graded-rollout-weighted pooled DiD is $+0.29$ pp. Both agents lose comparably—4.1 pp (GEMMA-4-31B-INSTRUCT) and 3.8 pp (QWEN-3.6-27B)—when the data switches from GEMMA-4-31B-

Language	Reviewer	Δ (pp)	95% CI	p-value
spa	GPT-OSS-120B	$+2.25$	$[-4.1, +8.6]$	0.49
por	GPT-OSS-120B	$+1.25$	$[-4.3, +6.8]$	0.66
fra	GPT-OSS-120B	$+1.50$	$[-4.3, +7.3]$	0.61
ita	GPT-OSS-120B	$+2.00$	$[-4.2, +8.2]$	0.53
deu	GPT-OSS-120B	$+1.00$	$[-5.0, +7.0]$	0.74
hin	GPT-OSS-120B	$+0.75$	$[-5.8, +7.3]$	0.82
cmn	GPT-OSS-120B	-0.75	$[-6.6, +5.1]$	0.80
jpn	GPT-OSS-120B	$+2.25$	$[-4.0, +8.5]$	0.48
tur	GPT-OSS-120B	$+2.25$	$[-4.3, +8.8]$	0.50
ind	GPT-OSS-120B	-2.25	$[-8.4, +3.9]$	0.47
spa	GEMMA-4-31B-INSTRUCT-self	$+1.25$	$[-5.0, +7.5]$	0.70
cmn	GEMMA-4-31B-INSTRUCT-self	$+1.75$	$[-4.2, +7.7]$	0.57
ind	GEMMA-4-31B-INSTRUCT-self	-1.75	$[-7.9, +4.4]$	0.58
spa	QWEN-3.6-27B	-0.75	$[-7.0, +5.5]$	0.81
cmn	QWEN-3.6-27B	$+0.50$	$[-5.4, +6.4]$	0.87
ind	QWEN-3.6-27B	-1.50	$[-7.7, +4.7]$	0.63

Table 4: Reviewer sweep: per-(language, reviewer) change in candidate-win rate Δ against the reviewer-free GEMMA-4-31B-INSTRUCT-only baseline. Every configuration's confidence interval crosses zero: no reviewer configuration produces a statistically detectable change on any language.

INSTRUCT-MT to QWEN-3.6-27B-MT, a symmetric *translator-quality* effect (GEMMA-4-31B-INSTRUCT is the stronger translator, Appendix B.1) rather than a family-of-origin advantage. To summarise across languages and to separate a genuine family-match effect from the overall difference in translator quality, we fit an ordinary least-squares (OLS) regression over the 40 cell-level observations (4 cells × 10 languages), weighting each cell equally. Writing y for a cell's completion-robust pass@1 (the pass rate among graded rollouts, $\text{pass}/(\text{pass} + \text{fail})$), we regress

$$y = \beta_0 + \beta_1 A + \beta_2 T + \beta_3 (A \cdot T) + \varepsilon, \tag{2}$$

where $A = 1$ for the GEMMA-4-31B-INSTRUCT agent (and 0 for QWEN-3.6-27B) and $T = 1$ when the evaluation data is QWEN-3.6-27B-translated (and 0 for GEMMA-4-31B-INSTRUCT-translated). The coefficients β_1 and β_2 are the two *main effects*—the average change in pass@1 from switching the agent, respectively the translator, with the other held fixed—and β_3, the coefficient on the product term, is the *family-match interaction*: the extra change when agent and translator switch together, i.e. how much more the GEMMA-4-31B-INSTRUCT agent gains from GEMMA-4-31B-INSTRUCT-translated data than the QWEN-3.6-27B agent does. Under this coding β_3 is the negative of the equally weighted cell-level DiD, so a GEMMA-4-31B-INSTRUCT home-field advantage would show up as $\beta_3 < 0$.

We estimate $\beta_3 = -0.37$ pp with a standard error (SE, the estimated sampling uncertainty of the coefficient) of 2.9 pp. Its magnitude differs slightly from the pooled $+0.29$ pp DiD because the regression weights each language cell equally, whereas the pooled estimate weights cells by their number of graded rollouts. The corresponding t-statistic $t = \beta_3/\text{SE} \approx 0.13$ is far below the $|t| \approx 2$ required for significance at the 5% level, so the interaction is statistically indistinguishable from zero. Re-fitting with an added covariate—each cell's measured translator quality, taken as the per-language candidate-win rate of Table 3—leaves β_3 essentially unchanged, so the (already negligible) family-match signal is not an artefact of translator quality varying across languages. Complementing these pooled and regression estimates, every one of the ten *per-language* DiD values has a 95% confidence interval—from a scenario-clustered bootstrap (resampling whole scenarios with replacement, $B = 1{,}000$ times, so that rollouts of the same scenario stay together)—that crosses zero under both the completion-robust and the raw metric (Figure 5); the largest per-language effect (tur, -3.3 pp on the robust metric) is *Qwen*-favoured, i.e. directionally opposite to a GEMMA-4-31B-INSTRUCT home-field advantage. We conclude the cross-lingual gap decomposition (Section 5) is not confounded by translator family.

Metric note. This ablation uses a distinct rollout-level estimand from the headline scenario-level metrics of Section 3. The GEMMA-4-31B-INSTRUCT×QWEN-3.6-27B-MT cells carry ≈ 13% non-graded rollouts (agent timeouts and no-verdict cases on long search trajectories at 262K-token

context), asymmetrically concentrated in that single arm. We therefore take completion-robust pass@1 = pass/(pass + fail) as the primary metric and report raw pass@1 as a robustness bound; the verdict is identical under both (pooled DiD $+0.29$ vs $+1.18$ pp), and the difference-in-differences subtracts out most of the completion-rate asymmetry.

	Evaluation data translated by	
Agent	GEMMA-4-31B-INSTRUCT-MT	QWEN-3.6-27B-MT
GEMMA-4-31B-INSTRUCT	31.2	27.2
QWEN-3.6-27B	32.7	28.9

Table 5: Family-match 2×2: graded-rollout-weighted pooled completion-robust pass@1 (%) over all ten target languages (1,920 nominal rollouts per cell). Both agents drop ≈ 4 pp on QWEN-3.6-27B-translated data; the difference-in-differences (GEMMA-4-31B-INSTRUCT home-field) is $+0.29$ pp. Rows are the evaluated agent; columns are the translator that produced the evaluation data.

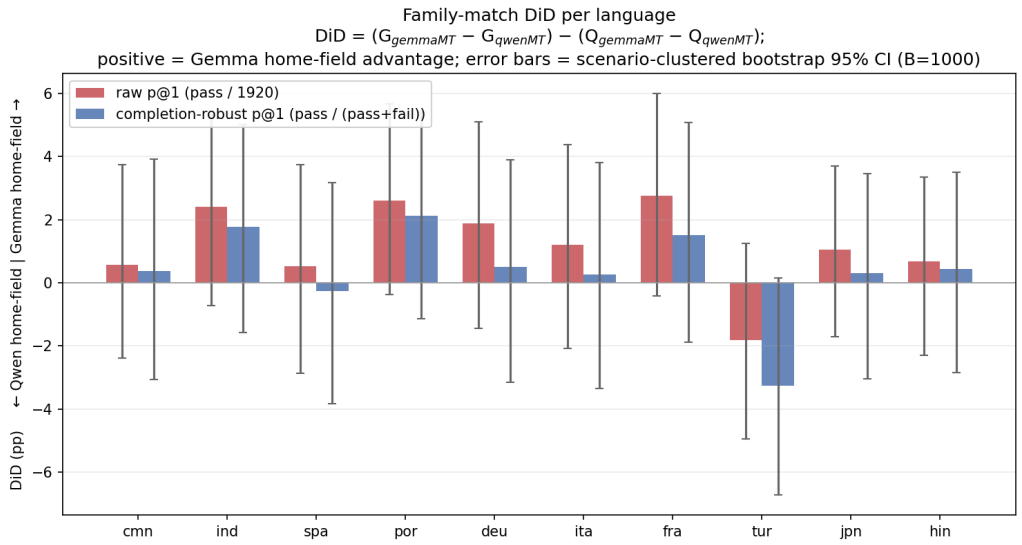

Figure 5: Per-language difference-in-differences (GEMMA-4-31B-INSTRUCT home-field advantage), raw and completion-robust, with 95% scenario-clustered bootstrap intervals ($B = 1{,}000$). All ten intervals cross zero under both metrics; positive values would indicate a GEMMA-4-31B-INSTRUCT family-match advantage. The one language with the largest magnitude (`tur`) is Qwen-favoured.

C JUDGE CALIBRATION

In this section we report additional details on the judge model and prompt ablation results described in Section 2.3, on the original GAIA2 human-annotated calibration set.

C.1 ENGLISH CALIBRATION

The four configurations cross the two judge models, the paper's reference LLAMA-3.3-70B-INSTRUCT and the new GPT-OSS-120B, with the two prompt sets, the upstream default and our localised overrides: J1 (LLAMA-3.3-70B-INSTRUCT, default), J2 (GPT-OSS-120B, default), J3 (GPT-OSS-120B, localised), and J4 (LLAMA-3.3-70B-INSTRUCT, localised). Table 6 reports overall Cohen's κ (Cohen, 1960) against the human-majority label on the $n = 414$ non-inconclusive traces, with bootstrap 95% confidence intervals. The remaining 48 of the 462 frozen traces have an inconclusive human-majority label: we exclude them from judge–human agreement statistics but retain them for the judge–judge comparisons below. Throughout this appendix, Δpass is the judge pass rate minus the fixed human-majority pass rate, reported as a proportion; negative values mean that the judge is stricter than the human label. J3 is the judge configuration we use when constructing the agent leaderboard in Section 4.

Config	Model	Prompts	Agreement	κ	95% CI	Δpass
J1	LLAMA-3.3-70B-INSTRUCT	default	0.879	0.714	[0.641, 0.781]	−0.111
J2	GPT-OSS-120B	default	0.882	0.720	[0.644, 0.787]	−0.109
J3	GPT-OSS-120B	localised	0.886	0.732	[0.661, 0.796]	−0.104
J4	LLAMA-3.3-70B-INSTRUCT	localised	0.882	0.720	[0.647, 0.789]	−0.109

Table 6: Overall judge agreement against the human-majority label on the $n = 414$ non-inconclusive traces of the recovered calibration set. All four κ values fall in $[0.714, 0.732]$ with heavily overlapping CIs: on English, neither the model swap nor prompt localisation materially changes agreement with humans. J3 matches the paper-standard reference J1.

Because all four configurations are scored on the same frozen traces, the two design choices can be isolated directly by judge-vs-judge agreement. Holding the model fixed at LLAMA-3.3-70B-INSTRUCT, localising the prompts (J1 $\rightarrow$ J4) flips the verdict on *one single* trace out of 462 ($\kappa = 0.994$, 95% CI $[0.978, 1.000]$); holding the prompts fixed at the localised set, swapping the model (J4 $\rightarrow$ J3) flips two ($\kappa = 0.987$, 95% CI $[0.967, 1.000]$); and the full swap from the paper's reference (J1 $\rightarrow$ J3) flips three ($\kappa = 0.981$, 95% CI $[0.957, 1.000]$), every one in the direction of J3 accepting a trajectory the reference rejected. On English, then, both the model choice and the prompt version are immaterial to judge–human agreement; the localisation payoff is realised in the target languages (Section 2.3).

Table 7 reports per-capability κ against the human-majority label alongside the human inter-annotator ceiling where multi-annotator coverage exists on this set.

Capability	Human agreement	J1	J2	J3	J4
adaptability	0.327	0.000	0.000	0.000	0.000
ambiguity	(single-annotation)	0.766	0.798	0.798	0.798
execution	(single-annotation)	0.595	0.619	0.642	0.595
search	(single-annotation)	1.000	0.979	1.000	1.000
time	0.104	0.000	0.000	0.000	0.000

Table 7: Per-capability Cohen's κ against the human-majority label under each judge configuration. The human inter-annotator agreement (IAA) column is Cohen's κ on the multi-annotator subset of the calibration set (adaptability $n = 83$, time $n = 45$); the other three capabilities are single-annotator on this set, so no human ceiling can be computed.

The $\kappa = 0$ entries on *adaptability* and *time* appear in *all four* configurations, not only under J4. On these capabilities the judge—like the human annotators—passes essentially no trace, so the label distribution has near-zero variance and Cohen's κ collapses to 0 regardless of raw agreement (which

exceeds 0.87 throughout). This is a property of the calibration corpus, not a judge weakness; see the failure-taxonomy discussion below.

Because the full-corpus κ of Table 6 is depressed both by these degenerate capabilities and, more broadly, by the deterministic-checker-dominated traces on which every configuration returns the same structural verdict, we recompute the 2×2 on the *LLM-adjudicated* slice—the $n = 153$ traces whose verdict actually invokes an LLM checker.[5] This is the *same* frozen slice used for the multilingual comparison (Table 10), so the English and cross-lingual numbers are directly comparable. On it (Table 8) every configuration rises to $\kappa \geq 0.928$, against 0.714–0.732 on the full corpus: the operational judge is strongest precisely where the LLM is exercised, and the full-corpus figures understate rather than overstate its reliability. The 2×2 ordering is preserved—localised $\geq$ default on each model, and GPT-OSS-120B $\geq$ LLAMA-3.3-70B-INSTRUCT under each prompt set—though all four CIs overlap, so the differences are directional, not significant.

Config	Model	Prompts	n	Agr.	κ	95% CI	Δpass
J1	LLAMA-3.3-70B-INSTRUCT	default	153	0.967	0.928	[0.859, 0.985]	−0.007
J2	GPT-OSS-120B	default	153	0.974	0.942	[0.881, 0.986]	0.000
J3	GPT-OSS-120B	localised	153	0.987	0.971	[0.923, 1.000]	+0.013
J4	LLAMA-3.3-70B-INSTRUCT	localised	153	0.974	0.942	[0.881, 0.986]	0.000

Table 8: The 2×2 recomputed on the LLM-adjudicated slice ($n = 153$; the same frozen traces as the multilingual Table 10). Removing the deterministic-checker-dominated and degenerate-κ capabilities lifts every configuration to $\kappa \geq 0.928$, versus 0.714–0.732 on the full corpus (Table 6). Localising the prompts on the fixed LLAMA-3.3-70B-INSTRUCT model (J1 → J4) still moves the verdict on essentially no trace ($\kappa = 0.986$ judge–judge on this slice). Agr. denotes raw agreement.

C.2 FAILURE TAXONOMY ON THE CALIBRATION SET

The calibration set is heavily deterministic-checker-dominated on three of the four evaluated capabilities: on adaptability, ambiguity and execution the verdict is decided by structured checks that never exercise the LLMaaJ component. The recovered calibration corpus also contains Time, as shown in Table 7, but Time is outside the current benchmark scope and is excluded from this operational and multilingual analysis. Table 9 reports the failure distribution over the four evaluated capabilities.

Capability	Success	Det. fail	LLM fail
adaptability	0	80/83 (96%)	3/83 (4%)
ambiguity	19	72/95 (76%)	4/95 (4%)
execution	29	64/94 (68%)	1/94 (1%)
search	54	0	43/97 (44%)

Table 9: Verdict provenance on the English calibration set, computed on the operational judge J3 (not the human labels). *Success* counts traces J3 passes; *Det. fail* counts J3 failures whose first failing check is deterministic (tool- or write-action-count, stuck-loop or timeout, or a deterministic content checker); *LLM fail* counts J3 failures whose first failing check is an LLM checker (`message`, `user_message`, `signature`, `tone`, or `content`). The three counts sum to n per capability. Only *search* exercises the LLMaaJ component meaningfully.

C.3 MULTILINGUAL ROBUSTNESS

Table 10 restricts the calibration set to the $n = 153$ traces where the LLM judge actually fires and reports per-language κ under J3 for eng and all ten target languages. This is the headline number of Section 2.3: it is the only slice on which translation of the natural-language content can plausibly move the verifier verdict. Because the English human-majority label is held fixed across languages

[5] A trace is LLM-adjudicated if it passes, or if its first failing turn is decided by one of the natural-language checkers (`message_checker`, `user_message_checker`, `signature_checker`, `tone_checker`, `content_checker`) or returns inconclusive; `time` is excluded, as in the multilingual slice.

(a faithful translation should not change the correct verdict), κ here isolates whether translation alone perturbs the judge.[6] The eight languages beyond eng, jpn and spa were scored on a freshly provisioned GPT-OSS-120B endpoint with identical `gpt-oss-120b` weights and `omnigaia_v2` prompts, differing only in serving host.

Language	n	κ	Judge pass	$\Delta\kappa$ vs eng
eng	153	0.971	0.667	—
cmn	153	0.971	0.654	0.000
jpn	153	0.957	0.660	−0.014
ita	153	0.942	0.654	−0.029
fra	153	0.928	0.647	−0.043
spa	153	0.914	0.641	−0.057
por	153	0.914	0.641	−0.057
hin	153	0.901	0.621	−0.070
deu	153	0.887	0.627	−0.084
ind	153	0.887	0.627	−0.084
tur	153	0.886	0.641	−0.085

Table 10: J3 agreement with human-majority labels on the LLM-adjudicated slice of the calibration set, for all eleven in-scope languages. The human pass rate on this slice is 0.654 in every language: the English human-majority label is translation-invariant, so κ isolates translation-induced verdict drift. Every $\Delta\kappa$ sits inside the pre-registered $\Delta\kappa \leq 0.10$ hazard threshold; the largest drop is tur at -0.085, and the non-Latin scripts (cmn, jpn, hin) rank among the strongest, so script family does not predict translation drift.

Per-capability multilingual κ is reported only for *search* (Table 11). We deliberately omit per-capability κ for adaptability, ambiguity and execution under translation: on those subsets the verdict is deterministic (see Table 9) and κ is language-invariant by construction, which would document a corpus property rather than judge quality.

Language	Agr.	κ	Δpass
eng	1.000	1.000	0.000
cmn	1.000	1.000	0.000
jpn	1.000	1.000	0.000
ita	0.979	0.958	0.000
fra	0.990	0.979	−0.010
spa	0.979	0.958	−0.021
por	0.990	0.979	−0.010
hin	0.990	0.979	−0.010
deu	0.979	0.958	−0.021
ind	0.979	0.958	0.000
tur	0.979	0.958	0.000

Table 11: J3 per-language agreement on the *search* capability ($n = 97$), the one capability with a substantial LLM-adjudicated share (43/97 in English). Agreement holds at $\kappa \geq 0.958$ across all eleven languages; eng, jpn and cmn are invariant, and the largest loss is two verdicts (spa, deu). Agr. denotes raw agreement.

As a same-family bias check on the operational judge, we additionally re-scored a held-out set of GPT-OSS-120B rollouts—used solely for this judge self-preference check, as GPT-OSS-120B is not part of the evaluated agent leaderboard—with both J3 (self) and the cross-family reference J1: the self−cross overall pass@1 difference is negative in all ten target languages ($\Delta \in [-0.025, -0.014]$), so GPT-OSS-120B shows no self-preference when adjudicating its own family's rollouts—if anything it is marginally stricter than the reference.

[6] In roughly 2% of translated traces (76 of 3,690) the GEMMA-4-31B-INSTRUCT translator regurgitated part of its own system prompt into the oracle user message, concentrated on the *search* capability whose user messages are longest. Manual review of the 72 affected traces that fall in the LLM-adjudicated slice found no verdict divergence from English on any of them, so the calibration figures reported here are unaffected.

D PER-LANGUAGE AND PER-CAPABILITY RESULTS

This appendix expands on the results reported in the main table (Table 1) and the heatmap (Figure 1) into the full per-agent, per-language, per-capability numbers.

D.1 PER-LANGUAGE BREAKDOWN

For each (agent, language, capability) cell, we report pass@1 and pass@3 with Wilson 95% confidence intervals (Wilson, 1927) and the all-three-consistency metric $\text{pass}_3^{\text{all}}$. Table 12 reports all four capabilities together, with the capabilities laid out across the columns.

Table 12: Full per-agent, per-language breakdown across the four capabilities. Each capability reports pass@1 / pass@3 / $\text{pass}_3^{\text{all}}$ (%); each cell contains 160 nominal scenarios. Lang denotes language; the compact headers @1, @3, and all denote pass@1, pass@3, and $\text{pass}_3^{\text{all}}$, respectively. Subscripts and superscripts on pass@1/pass@3 are Wilson 95% CI bounds.

		Exec.			Search			Adapt.			Ambig.		
Agent	Lang	@1	@3	all	@1	@3	all	@1	@3	all	@1	@3	all
CLAUDE-4.7-OPUS	eng	79.4^{83}_{76}	88.1^{92}_{82}	69.4	84.0^{87}_{80}	95.6^{98}_{91}	67.7	63.9^{68}_{59}	75.6^{82}_{68}	49.4	58.0^{62}_{53}	71.6^{78}_{64}	43.2
	spa	70.4^{74}_{66}	84.4^{89}_{78}	54.4	74.6^{78}_{70}	89.3^{93}_{84}	57.9	61.8^{66}_{57}	75.6^{82}_{68}	43.8	49.3^{54}_{45}	68.8^{76}_{61}	29.9
	deu	65.8^{70}_{61}	79.4^{85}_{72}	48.1	58.4^{63}_{54}	78.0^{84}_{71}	34.6	58.0^{63}_{53}	70.0^{77}_{62}	47.5	42.2^{47}_{38}	58.3^{66}_{50}	28.2
	fra	66.7^{71}_{62}	83.1^{88}_{77}	47.5	65.1^{70}_{60}	81.1^{86}_{74}	50.3	60.5^{65}_{56}	72.5^{79}_{65}	46.2	42.0^{47}_{38}	61.1^{68}_{53}	24.8
	ita	59.8^{65}_{54}	76.2^{82}_{69}	41.2	69.0^{73}_{64}	81.0^{86}_{74}	55.7	62.2^{68}_{56}	74.4^{81}_{67}	48.8	35.3^{40}_{31}	54.1^{66}_{46}	18.5
	por	64.9^{70}_{60}	81.9^{87}_{75}	46.2	70.6^{75}_{66}	84.9^{90}_{79}	52.2	54.9^{59}_{50}	73.8^{80}_{66}	35.0	41.3^{46}_{37}	60.3^{68}_{52}	23.7
	ind	69.5^{73}_{65}	83.1^{88}_{77}	55.6	69.4^{73}_{65}	86.2^{91}_{80}	50.3	61.6^{66}_{57}	73.8^{80}_{66}	45.6	44.0^{49}_{40}	60.9^{68}_{53}	26.9
	tur	61.9^{67}_{57}	77.5^{83}_{70}	45.6	68.3^{72}_{64}	83.0^{88}_{76}	52.2	52.0^{56}_{48}	72.5^{79}_{65}	31.9	42.4^{47}_{38}	66.7^{74}_{59}	19.2
	cmn	62.4^{67}_{58}	75.0^{81}_{68}	49.4	66.0^{70}_{62}	81.8^{87}_{75}	47.2	56.7^{61}_{52}	73.1^{79}_{66}	38.1	40.5^{45}_{36}	55.8^{63}_{48}	28.8
	jpn	56.1^{61}_{52}	70.0^{77}_{62}	41.2	65.3^{69}_{61}	83.0^{88}_{76}	47.8	55.3^{60}_{50}	73.0^{79}_{66}	33.3	35.9^{40}_{32}	51.9^{60}_{44}	19.0
	hin	53.7^{58}_{49}	77.5^{83}_{70}	30.0	63.8^{68}_{59}	81.8^{87}_{75}	44.0	54.8^{60}_{50}	71.9^{78}_{64}	36.2	40.5^{45}_{36}	59.2^{67}_{51}	21.0
GPT-5.4	eng	63.0^{67}_{59}	81.2^{87}_{74}	44.4	82.1^{85}_{78}	92.4^{96}_{87}	70.1	39.2^{44}_{35}	58.8^{66}_{51}	21.9	25.6^{30}_{22}	41.4^{49}_{34}	12.7
	spa	51.9^{57}_{47}	69.4^{76}_{62}	33.1	74.8^{78}_{71}	87.3^{92}_{81}	61.8	34.8^{39}_{31}	51.2^{59}_{44}	18.1	19.0^{23}_{16}	33.3^{41}_{26}	8.2
	deu	52.5^{57}_{48}	68.8^{75}_{61}	32.5	72.1^{76}_{68}	85.4^{90}_{79}	56.1	36.0^{40}_{32}	53.8^{61}_{46}	19.4	21.4^{25}_{18}	35.4^{43}_{28}	8.9
	fra	49.9^{54}_{45}	65.0^{72}_{57}	31.9	75.1^{79}_{71}	88.6^{93}_{83}	61.4	35.6^{40}_{31}	52.5^{60}_{45}	17.5	21.1^{25}_{18}	34.4^{42}_{27}	9.6
	ita	49.8^{54}_{45}	70.0^{77}_{62}	28.7	73.6^{77}_{69}	86.7^{91}_{81}	57.6	38.5^{43}_{34}	57.5^{70}_{50}	18.8	17.3^{21}_{14}	29.1^{37}_{23}	7.6
	por	53.7^{58}_{49}	73.8^{80}_{66}	34.4	71.4^{75}_{67}	85.3^{90}_{79}	56.4	34.5^{39}_{30}	51.2^{59}_{44}	17.5	19.6^{23}_{16}	33.1^{41}_{26}	8.9
	ind	52.8^{58}_{48}	70.6^{77}_{63}	30.6	71.9^{76}_{68}	84.6^{89}_{78}	55.8	35.0^{39}_{31}	54.4^{67}_{47}	16.9	19.7^{24}_{16}	32.5^{40}_{26}	9.6
	tur	53.1^{58}_{49}	73.8^{80}_{66}	35.0	66.8^{71}_{62}	78.5^{84}_{71}	52.5	35.6^{40}_{31}	52.5^{60}_{45}	19.4	20.9^{25}_{17}	34.6^{42}_{28}	10.7
	cmn	41.8^{46}_{37}	57.2^{65}_{49}	24.5	65.5^{70}_{61}	80.3^{86}_{73}	48.4	33.3^{38}_{29}	48.8^{56}_{41}	17.5	18.2^{22}_{15}	29.9^{38}_{23}	10.2
	jpn	40.5^{46}_{35}	58.1^{65}_{50}	24.4	61.1^{67}_{57}	77.6^{83}_{70}	42.3	24.9^{29}_{21}	43.1^{51}_{36}	10.0	15.6^{19}_{13}	26.8^{34}_{20}	8.3
	hin	45.4^{50}_{41}	64.2^{71}_{56}	24.5	66.0^{70}_{62}	80.8^{86}_{74}	53.2	31.2^{36}_{27}	49.4^{57}_{42}	14.4	16.4^{20}_{13}	29.1^{37}_{23}	5.7
GEMINI-3.1-PRO	eng	48.1^{53}_{42}	66.2^{73}_{59}	28.8	68.3^{72}_{64}	86.9^{91}_{81}	47.1	31.2^{36}_{27}	51.2^{59}_{44}	12.5	26.9^{31}_{23}	44.4^{52}_{37}	12.0
	spa	37.9^{43}_{34}	56.2^{64}_{49}	20.0	62.4^{68}_{58}	80.5^{86}_{74}	43.4	29.6^{34}_{26}	46.2^{54}_{39}	13.1	21.2^{25}_{18}	34.8^{43}_{28}	10.1
	deu	38.4^{43}_{34}	59.4^{67}_{52}	19.4	58.8^{63}_{54}	79.7^{85}_{73}	38.6	31.2^{36}_{27}	50.6^{58}_{43}	12.5	20.8^{25}_{17}	37.3^{45}_{30}	6.3
	fra	40.1^{45}_{36}	55.6^{68}_{48}	25.6	60.0^{66}_{56}	77.2^{83}_{70}	43.0	31.0^{37}_{27}	51.2^{59}_{44}	11.9	18.8^{23}_{15}	30.2^{38}_{24}	8.2
	ita	36.0^{40}_{32}	53.8^{61}_{46}	18.1	60.1^{64}_{56}	79.1^{85}_{72}	37.3	29.8^{34}_{26}	51.2^{59}_{44}	10.0	18.9^{23}_{16}	32.1^{40}_{25}	8.2
	por	39.2^{44}_{35}	60.0^{67}_{52}	20.6	59.7^{64}_{55}	77.4^{83}_{70}	40.9	30.8^{35}_{27}	53.1^{61}_{45}	11.2	18.7^{22}_{15}	31.6^{39}_{25}	8.9
	ind	38.2^{43}_{34}	58.8^{66}_{51}	18.8	62.2^{68}_{58}	80.5^{86}_{74}	42.1	31.5^{37}_{27}	48.8^{56}_{41}	16.9	16.5^{20}_{13}	28.9^{38}_{22}	7.5
	tur	37.7^{42}_{33}	54.4^{62}_{47}	21.9	53.2^{58}_{49}	75.8^{82}_{69}	31.2	25.5^{30}_{22}	46.2^{54}_{39}	8.8	17.0^{21}_{14}	28.5^{36}_{22}	9.5
	cmn	41.8^{47}_{37}	60.6^{68}_{53}	23.1	56.9^{62}_{52}	77.2^{83}_{70}	34.8	28.6^{33}_{25}	50.6^{68}_{43}	7.5	14.9^{18}_{12}	24.1^{31}_{18}	8.2
	jpn	36.3^{41}_{32}	53.1^{61}_{45}	19.4	54.9^{59}_{50}	74.8^{81}_{68}	34.6	24.8^{29}_{21}	43.1^{51}_{36}	8.1	16.5^{20}_{13}	26.6^{34}_{20}	7.0
	hin	31.9^{36}_{28}	49.4^{57}_{42}	16.9	58.4^{63}_{54}	79.0^{85}_{72}	36.9	25.0^{29}_{21}	47.5^{55}_{40}	8.1	14.3^{18}_{11}	24.5^{32}_{18}	6.3
KIMI-2.6	eng	51.7^{56}_{47}	73.8^{80}_{66}	28.7	81.5^{85}_{78}	94.3^{97}_{90}	63.9	35.7^{40}_{32}	55.0^{63}_{47}	18.8	25.5^{30}_{22}	42.8^{51}_{35}	9.4
	spa	36.9^{41}_{33}	57.5^{65}_{50}	18.1	73.4^{77}_{69}	91.2^{95}_{86}	53.5	31.5^{36}_{28}	48.1^{56}_{41}	17.5	20.5^{24}_{17}	36.7^{44}_{30}	8.2
	deu	37.1^{42}_{33}	55.6^{63}_{48}	19.4	64.9^{69}_{60}	82.3^{87}_{76}	43.0	31.4^{36}_{27}	48.1^{56}_{41}	15.6	17.3^{21}_{14}	28.5^{36}_{22}	7.6
	fra	36.9^{41}_{33}	56.2^{64}_{49}	18.8	70.9^{75}_{67}	85.5^{90}_{79}	51.6	31.5^{36}_{28}	46.9^{55}_{39}	17.5	21.3^{25}_{18}	35.0^{43}_{28}	7.6
	ita	38.3^{43}_{34}	58.1^{65}_{50}	18.1	69.9^{74}_{66}	87.3^{92}_{81}	48.7	31.9^{36}_{28}	48.8^{56}_{41}	15.0	19.5^{23}_{16}	30.6^{38}_{24}	10.0
	por	38.5^{43}_{34}	59.4^{67}_{52}	20.6	72.3^{76}_{68}	88.1^{92}_{82}	53.5	26.1^{30}_{22}	42.5^{50}_{35}	13.8	16.8^{20}_{14}	28.7^{36}_{22}	6.9
	ind	33.8^{38}_{30}	50.6^{63}_{43}	16.2	70.7^{75}_{66}	89.9^{94}_{84}	46.2	25.5^{30}_{22}	41.2^{49}_{34}	10.6	16.1^{20}_{13}	29.4^{37}_{23}	5.0
	tur	36.7^{41}_{32}	62.5^{70}_{55}	13.8	63.1^{67}_{59}	79.0^{85}_{72}	44.6	26.7^{31}_{23}	45.6^{53}_{38}	9.4	17.2^{21}_{14}	30.8^{38}_{24}	5.0
	cmn	35.0^{39}_{31}	53.1^{61}_{45}	15.6	64.6^{69}_{60}	83.0^{88}_{76}	45.3	26.3^{30}_{23}	40.0^{48}_{33}	13.8	21.0^{25}_{18}	34.0^{42}_{27}	8.8
	jpn	35.4^{40}_{31}	53.8^{61}_{46}	16.2	63.4^{68}_{59}	79.6^{85}_{73}	43.9	25.9^{30}_{22}	43.1^{51}_{36}	10.6	16.7^{20}_{14}	30.4^{38}_{24}	5.7
	hin	33.8^{38}_{30}	58.1^{65}_{50}	15.0	61.9^{66}_{57}	82.4^{88}_{76}	39.0	24.0^{28}_{20}	40.3^{48}_{33}	10.1	14.1^{18}_{11}	26.9^{34}_{21}	5.6
GEMMA-4-31B-INSTRUCT	eng	50.3^{55}_{46}	67.5^{74}_{60}	34.4	70.9^{75}_{67}	85.6^{90}_{79}	50.0	45.2^{50}_{41}	63.1^{70}_{55}	26.2	23.8^{28}_{20}	36.9^{45}_{30}	10.6
	spa	41.9^{46}_{38}	56.9^{64}_{49}	25.6	63.6^{68}_{59}	78.1^{84}_{71}	41.2	39.0^{43}_{35}	57.5^{65}_{50}	23.1	18.3^{22}_{15}	28.1^{36}_{22}	8.8
	deu	45.6^{50}_{41}	58.8^{66}_{51}	32.5	59.1^{65}_{55}	76.2^{82}_{69}	40.0	37.3^{42}_{33}	57.5^{65}_{50}	19.4	18.5^{22}_{15}	31.9^{39}_{25}	6.9
	fra	42.1^{47}_{38}	59.4^{67}_{52}	23.8	62.2^{67}_{58}	75.6^{82}_{68}	48.1	41.2^{46}_{37}	56.2^{64}_{49}	27.5	18.9^{23}_{16}	31.2^{39}_{25}	8.8
	ita	43.4^{48}_{39}	60.0^{67}_{52}	28.8	63.6^{68}_{59}	76.2^{82}_{69}	42.5	41.8^{46}_{38}	58.1^{65}_{50}	26.9	16.8^{20}_{14}	28.1^{36}_{22}	8.1

(Table 12 continued)

Agent	Lang	Exec. @1	Exec. @3	Exec. all	Search @1	Search @3	Search all	Adapt. @1	Adapt. @3	Adapt. all	Ambig. @1	Ambig. @3	Ambig. all
	por	47.3^{52}_{43}	62.5^{70}_{55}	31.2	60.8^{65}_{56}	71.2^{78}_{64}	40.6	35.5^{40}_{31}	50.6^{58}_{43}	20.0	18.6^{22}_{15}	28.1^{36}_{22}	8.8
	ind	45.2^{50}_{41}	61.2^{68}_{54}	27.5	60.2^{65}_{56}	75.0^{81}_{68}	37.5	40.4^{45}_{36}	59.4^{67}_{52}	20.6	18.5^{22}_{15}	31.2^{39}_{25}	8.1
	tur	40.5^{45}_{36}	56.9^{64}_{49}	25.0	52.5^{57}_{48}	65.0^{72}_{57}	36.9	35.8^{40}_{32}	50.6^{58}_{43}	20.6	18.3^{22}_{15}	30.6^{38}_{24}	6.9
	cmn	35.2^{40}_{31}	49.4^{57}_{42}	20.0	46.3^{51}_{42}	63.1^{70}_{55}	26.2	36.2^{41}_{32}	48.1^{56}_{41}	23.1	18.4^{22}_{15}	25.6^{33}_{19}	10.0
	jpn	31.6^{36}_{28}	41.9^{50}_{35}	21.2	42.3^{47}_{38}	59.4^{67}_{52}	21.2	30.5^{35}_{27}	47.5^{55}_{40}	11.9	13.1^{16}_{10}	20.0^{27}_{15}	7.5
	hin	28.5^{33}_{25}	45.0^{53}_{37}	14.4	39.9^{44}_{35}	55.0^{63}_{47}	20.6	33.7^{38}_{30}	48.8^{56}_{41}	18.1	13.0^{16}_{10}	20.0^{27}_{15}	5.6
QWEN-3.6-35B-A3B	eng	49.0^{53}_{45}	68.8^{75}_{61}	28.1	63.2^{67}_{59}	86.2^{91}_{80}	34.4	32.1^{36}_{28}	48.1^{56}_{41}	16.2	11.7^{15}_{9}	23.1^{30}_{17}	3.1
	spa	26.5^{31}_{23}	46.2^{54}_{39}	10.6	41.4^{46}_{37}	69.4^{76}_{62}	16.9	22.3^{26}_{19}	37.5^{45}_{30}	5.6	6.4^{9}_{5}	14.4^{21}_{10}	1.2
	deu	28.0^{32}_{24}	48.1^{56}_{41}	11.2	36.8^{41}_{33}	65.6^{73}_{58}	10.0	17.6^{21}_{14}	36.9^{45}_{30}	2.5	6.8^{11}_{5}	11.9^{18}_{8}	3.1
	fra	27.8^{32}_{24}	44.4^{52}_{37}	10.6	44.7^{49}_{40}	71.2^{78}_{64}	16.9	20.3^{24}_{17}	35.0^{43}_{28}	6.2	8.4^{11}_{6}	15.0^{21}_{10}	2.5
	ita	28.5^{33}_{25}	48.1^{56}_{41}	11.9	40.9^{45}_{36}	66.9^{74}_{59}	13.1	19.4^{23}_{16}	33.1^{41}_{26}	6.9	8.2^{11}_{6}	13.8^{20}_{9}	4.4
	por	28.5^{33}_{25}	47.5^{55}_{40}	10.0	40.4^{45}_{36}	68.8^{75}_{61}	11.9	19.9^{24}_{17}	32.5^{40}_{26}	8.8	8.5^{11}_{6}	15.6^{22}_{11}	1.9
	ind	25.7^{30}_{22}	42.5^{50}_{35}	9.4	40.2^{45}_{36}	66.9^{74}_{59}	13.8	21.3^{25}_{18}	40.6^{48}_{33}	4.4	7.2^{10}_{5}	13.1^{19}_{9}	1.9
	tur	22.7^{27}_{19}	40.6^{48}_{33}	8.1	35.7^{40}_{31}	63.1^{70}_{55}	11.9	15.8^{19}_{13}	29.4^{37}_{23}	3.8	7.2^{10}_{5}	14.4^{21}_{10}	1.2
	cmn	27.4^{32}_{24}	41.9^{50}_{35}	11.9	38.7^{43}_{34}	63.1^{70}_{55}	13.1	17.5^{21}_{14}	34.4^{42}_{27}	3.1	8.8^{12}_{7}	13.1^{19}_{9}	4.4
	jpn	17.8^{22}_{15}	30.0^{38}_{23}	7.5	35.7^{40}_{31}	61.9^{69}_{54}	6.9	12.3^{16}_{10}	23.1^{30}_{17}	5.0	6.7^{9}_{5}	13.8^{20}_{9}	2.5
	hin	15.9^{19}_{13}	28.1^{36}_{22}	5.0	31.6^{36}_{27}	58.8^{66}_{51}	8.1	8.1^{11}_{6}	15.6^{22}_{11}	1.9	5.3^{8}_{4}	10.6^{16}_{7}	0.0
QWEN-3.6-27B	eng	54.8^{59}_{50}	77.5^{83}_{70}	26.9	58.4^{63}_{54}	79.4^{85}_{72}	31.2	39.6^{44}_{35}	61.2^{68}_{54}	15.0	20.0^{24}_{17}	30.6^{38}_{24}	6.2
	spa	38.3^{43}_{34}	59.4^{67}_{52}	15.0	49.9^{55}_{45}	76.2^{82}_{69}	20.6	38.8^{43}_{35}	60.0^{67}_{54}	15.6	15.5^{19}_{12}	25.6^{33}_{19}	4.4
	deu	39.0^{44}_{35}	59.4^{67}_{52}	18.1	46.3^{51}_{42}	66.9^{74}_{59}	20.0	36.5^{41}_{32}	54.4^{62}_{47}	14.4	14.3^{18}_{11}	24.4^{32}_{18}	5.0
	fra	40.3^{45}_{36}	58.8^{66}_{51}	16.2	47.5^{52}_{43}	73.8^{80}_{66}	18.8	39.3^{44}_{35}	59.4^{67}_{52}	16.2	13.1^{17}_{10}	22.5^{30}_{17}	5.6
	ita	40.9^{45}_{36}	60.0^{67}_{52}	18.1	50.9^{56}_{46}	74.4^{81}_{67}	22.5	36.9^{41}_{33}	56.2^{64}_{49}	17.5	16.0^{20}_{13}	29.4^{37}_{23}	5.6
	por	39.5^{44}_{35}	58.8^{66}_{51}	18.1	50.5^{55}_{46}	69.4^{76}_{62}	21.9	32.4^{37}_{28}	50.6^{58}_{43}	10.6	15.5^{19}_{12}	27.5^{35}_{21}	5.0
	ind	39.4^{44}_{35}	61.9^{69}_{54}	14.4	48.3^{53}_{44}	70.6^{77}_{63}	18.1	32.3^{37}_{28}	51.9^{59}_{44}	11.9	12.4^{16}_{10}	21.9^{29}_{16}	3.1
	tur	33.3^{38}_{29}	54.4^{62}_{47}	13.1	44.6^{49}_{40}	64.4^{71}_{57}	18.8	32.6^{37}_{29}	53.1^{61}_{45}	13.1	15.9^{20}_{13}	28.8^{36}_{22}	5.0
	cmn	33.3^{38}_{29}	53.8^{61}_{46}	15.0	43.5^{48}_{39}	68.8^{75}_{61}	13.1	33.5^{38}_{29}	53.1^{61}_{45}	11.2	15.8^{19}_{13}	27.5^{35}_{21}	3.1
	jpn	27.5^{32}_{24}	43.1^{51}_{36}	11.2	40.6^{45}_{36}	66.2^{73}_{59}	11.2	23.4^{28}_{20}	38.8^{46}_{32}	8.1	11.6^{15}_{9}	19.4^{26}_{14}	4.4
	hin	22.8^{27}_{19}	42.5^{50}_{35}	3.1	37.2^{42}_{33}	63.8^{71}_{56}	10.0	24.3^{28}_{21}	41.9^{50}_{35}	7.5	9.5^{13}_{7}	16.9^{23}_{12}	3.8

E QWEN-3.5 SCALE LADDER: PER-LANGUAGE CELLS

We vary *scale* within a single family to ask how agentic competence and the cross-lingual gap of Section 4.2 move with model size. We evaluate four QWEN-3.5 sizes: one dense (27B) and three mixture-of-experts (MoE) models (35B-A3B, 122B-A10B, and 397B-A17B, with 3B, 10B, and 17B active parameters, respectively)[7]. We run the full benchmark suite across eleven in-scope languages and four capabilities under the same experimental setup. This scaling study uses a different model generation from QWEN-3.6-35B-A3B, so it complements rather than extends Table 1.

Model	Act.	Exec.	Search	Adapt.	Ambig.	Avg.
35B-A3B	3B	10.1	22.0	5.3	4.4	10.5
122B-A10B	10B	9.3	30.1	8.7	3.5	12.9
27B	27B	**21.4**	30.6	12.7	**7.6**	18.1
397B-A17B	17B	20.7	**34.3**	**17.4**	7.5	**20.0**

Table 13: QWEN-3.5 scale ladder: pass@1 (%) pooled over the eleven languages, per capability and averaged (`time` excluded), with the active-parameter count per size. Rows are ordered by average score; best per column in **bold**. Act., Exec., Adapt., Ambig., and Avg. denote active parameters, execution, adaptability, ambiguity, and average, respectively. Same protocol as the headline (pass@1, `thinking_effort`=low, operational judge J3); each cell uses a denominator of 160 scenarios per language-capability pair. Per-language cells are in Table 14.

Scaling is monotonic, but a dense mid-size rivals the largest MoE. Average pass@1 rises monotonically with the ladder: $10.5 \rightarrow 12.9 \rightarrow 18.1 \rightarrow 20.0$. 397B-A17B is the strongest size on average and on search and adaptability. But the dense 27B actually edges the $6\times$-larger 397B-A17B on execution (21.4 vs 20.7) and ambiguity (7.6 vs 7.5), and clears 122B-A10B comfortably.

[7]We omit checkpoints smaller than 9B, for which the agent harness does not reliably elicit tool calls, yielding near-zero valid trajectories.

Aggregate capability thus tracks the *active* parameter count, and the dense mid-size model is fully competitive with the largest sparse one per active parameter.

The capability ordering is stable. The canonical ordering of search > execution > adaptability > ambiguity holds across sizes, with ambiguity hardest throughout; the one exception is 122B-A10B, whose execution collapses to the level of its adaptability. Scale lifts the whole profile rather than re-ordering which capabilities are easy.

The cross-lingual gap does not close with scale. English is the ceiling at every size, and the gap to the target-language mean *widens* as models grow: English minus the mean of the other ten languages is +8.1, +7.5, +10.4, and +12.7 pp for 35B-A3B →397B-A17B. `hin` and `jpn` are the weakest languages at every size (e.g. 397B-A17B: `eng` 31.6 vs `hin` 9.2, `jpn` 10.0), mirroring the script-family fragility of Section 4.2. Scaling the model therefore raises the multilingual floor but leaves the *shape* of the cross-lingual gap intact, so the gap identified in this work is not one that a larger same-family checkpoint closes on its own.

Table 14 expands the pooled scale ladder in Table 13 into per-language pass@1 for the four QWEN-3.5 sizes, averaged over the four in-scope capabilities. Each cell pools $4 \times 160 = 640$ scenarios per (size, language). English is the strongest column at every size and `hin` the weakest, with `jpn` close behind; the English-minus-target gap is visible as the spread between the first column and the rest and *widens* with scale.

Rung	`eng`	`cmn`	`deu`	`fra`	`hin`	`ind`	`ita`	`jpn`	`por`	`spa`	`tur`
35B-A3B	17.8	7.7	11.6	12.5	3.0	14.7	10.2	4.7	11.2	13.9	7.8
122B-A10B	19.7	11.9	13.1	14.5	6.6	14.8	14.2	7.0	14.4	15.5	9.8
27B	27.5	15.9	18.9	21.4	7.3	19.2	20.3	**12.5**	19.4	22.5	13.6
397B-A17B	**31.6**	**16.2**	**21.1**	**22.5**	**9.2**	**21.6**	**23.9**	10.0	**23.0**	**23.1**	**17.7**

Table 14: QWEN-3.5 scale ladder, per-language pass@1 (%) averaged over the four in-scope capabilities (execution, search, adaptability, and ambiguity), pooling $4 \times 160 = 640$ scenarios per (size, language). Best per column in **bold**: 397B-A17B leads every language except `jpn` (won by the dense 27B). English is the ceiling and `hin`/`jpn` the floor at every size. Protocol and caveats are as in Table 13.

Reasoning budget (thinking effort). Increasing the model's thinking effort from low to high does not improve agentic pass@1 on this benchmark. On the largest 397B-A17B size, a high-effort rerun matches the low-effort ladder to within about 2 pp per capability and is, if anything, marginally *lower* (adaptability 17.0 vs 17.4; ambiguity 5.2 vs 7.5; execution tracking similarly). We therefore report the ladder at `thinking_effort`=low, matching the headline open-weight configuration.

F FULL PER-CAPABILITY, BEHAVIOURAL, AND VERIFIER-CHECK ANALYSIS

This appendix expands on the three analyses summarised in Section 4.2: the per-capability breakdown of the headline leaderboard, a per-scenario behavioural comparison of English and target-language rollouts, and a decomposition of the regression by which verifier check first fails.

F.1 PER-CAPABILITY BREAKDOWN

We break down the numbers of Table 1 into the four capabilities. For each (capability, language, agent) combination we report pass@3 with 95% confidence intervals (CIs) and test the English–target-language difference with a paired McNemar test (McNemar, 1947) on the matched scenario pairs, per capability. Figure 1 gives the full capability-by-language view at a glance, while the exact numbers, Wilson CIs, and per-cell McNemar p-values are in Appendix D.

Two patterns are visible: performance is strongest in English for nearly every (agent, capability) combination, and the drop concentrates on the three non-Latin-script languages (`cmn`, `jpn`, `hin`). These languages use non-Latin scripts, and we find that part of the reason is the agents not carefully

considering the various forms *entities* may take when calling tools (see Appendix J for details), mirroring cross-script mistakes in non-agentic tasks (Bandarkar et al., 2026). The proprietary agents (CLAUDE-4.7-OPUS, GPT-5.4) show the most cross-lingually robust behaviour, while the open-weight agents perform significantly below English on almost every target language.

Furthermore, the capability ordering appears stable across languages, with pass@3 numbers decreasing from search through execution and adaptability to ambiguity, so translation reshapes the *level* of each capability without re-ordering which capabilities are easy. Search is both the strongest capability and among the more robust, with KIMI-2.6 achieving the highest target-language search of any agent (84.7, edging CLAUDE-4.7-OPUS's 83.2 and GPT-5.4's 83.1). The proprietary agents are most cross-lingually robust on *adaptability*: CLAUDE-4.7-OPUS and GEMINI-3.1-PRO retain it almost intact in the target languages (means of 73.0 and 48.9, against English scores of 75.6 and 51.2, leading to gaps of 2.6 and 2.3 pp), so their pooled loss is carried instead by execution and, especially, ambiguity, which is the hardest capability everywhere and the one on which every agent scores lowest in the target languages.

Looking at the pass@1, pass@3 and $\text{pass}_3^{\text{all}}$ numbers (reported in Appendix D), we observe that for most agents extra rollouts do not recover the loss and the same target-language scenarios fail every time. CLAUDE-4.7-OPUS is the exception: its gap narrows from 14.0 pp at pass@1 to 9.1 pp at pass@3 yet is widest at $\text{pass}_3^{\text{all}}$ (17.5 pp), so in the target languages it still reaches the solution within three attempts, only far less consistently. That target-language pass@3 stays high (e.g. $\approx 74\%$ for CLAUDE-4.7-OPUS, 58% for GPT-5.4) shows these scenarios remain largely solvable in the target languages rather than the benchmark saturating; how much of the residual difficulty reflects genuine agent limitations versus translation artefacts is what we disentangle in the attribution analysis (Section 5).

Since GEMMA-4-31B-INSTRUCT serves as both the model in the translation pipeline (Section 2.2) and one of the evaluated agents, we verify that this leaderboard is not distorted by a *family-match* advantage, i.e. an agent scoring higher on data translated by its own model family. We run a Gemma–Qwen agent–translator ablation over all ten target languages, and find the agent gap invariant to which family translated the evaluation data, with both agents losing comparably on the weaker Qwen translations; the full design and per-language and per-capability results are in Appendix B.3. Additional scale ablation is reported in Appendix E.

F.2 BEHAVIOURAL SIGNATURES

Beyond pass rates, per-scenario signals—step count and tool-call composition—reveal which behavioural modes shift under translation: inflated step counts and retry loops would indicate an agent struggling with a partially understood instruction, whereas truncated planning or a collapsed tool-call vocabulary would suggest it giving up early rather than exploring. We analyse two proprietary agents (CLAUDE-4.7-OPUS, GPT-5.4) and the strongest open-weight agent (KIMI-2.6), pairing each target-language rollout with the matching English rollout.

Volume does not explain the regression. Only GPT-5.4 does systematically more work in the target languages ($+21.6\%$ tool calls, in *every* target language, up to $+41\%$ on `hin`); CLAUDE-4.7-OPUS and KIMI-2.6 issue slightly *fewer* (-7.6% on average, mixed in sign per language). Neither direction tracks accuracy in the way a flailing account would predict: for GPT-5.4 the scenarios that lose accuracy inflate their tool count no more than those that hold it ($+4.8$ vs $+6.5$ calls), so volume and success move independently, and the one agent whose tool count does track accuracy is CLAUDE-4.7-OPUS. Running the *opposite* way, its regressing scenarios issue -11.0 calls against $+3.6$ for the rest, a "give up early" signature.

Strategy shifts far more than a pooled average suggests. We quantify *which* tools each agent reaches for with the Jensen–Shannon divergence (JSD) between its English and target-language tool-choice distributions. Pooled over a cohort the shift looks negligible (0.010/0.025/0.020 bits for CLAUDE-4.7-OPUS/GPT-5.4/KIMI-2.6) because opposing per-scenario shifts cancel; measured per matched scenario it is several times larger (0.087, 0.109, 0.116), and larger still within matched thirds of each trajectory (0.170, 0.197, 0.216). The pooled figure therefore badly understates how much each agent re-plans in the target languages, and per scenario the three agents look much more

alike: GPT-5.4's apparent $2.5\times$ lead over CLAUDE-4.7-OPUS shrinks to about $1.3\times$, and KIMI-2.6 re-plans as much as either proprietary agent.

The shift is from writes to reads. Its direction is consistent across agents and languages: in the target languages, a larger fraction of the tool budget goes to *exploratory reads* and a smaller fraction to *state-changing writes*. Retrieval and browsing operations grow most (product search and catalogue listing $+3$–6 pp for every agent), while the write-action share of all tool calls falls from 32.1% to 29.9% for CLAUDE-4.7-OPUS, 30.5% to 28.1% for GPT-5.4 and 29.3% to 28.0% for KIMI-2.6 (-1.3 to -2.5 pp). In other words, in the target languages all three agents forage longer and commit later and less decisively: the behavioural counterpart to the tool-count-mismatch failures of Section 4.2, where the wrong *number* of side-effectful calls is the dominant target-language failure signature.

F.3 VERIFIER DECOMPOSITION

To identify *which* verifier checks drive the regression, we label each failed rollout by the first check it fails (quantitative, e.g. a mismatched count of retrieved emails; categorical, e.g. a mismatched attendee-set; textual, e.g. message content rejected by the LLMaaJ; overall tool-count mismatch; or timeout) and report each class's *absolute* failure rate, with all rollouts in the denominator: the share-of-failures view is flat across languages and hides the effect, whereas the absolute view localises it.

The regression is *not* carried by quantitative or categorical reasoning. From English to the target-language average, every quantitative- and categorical-checker delta falls within ± 0.35 pp for all three agents, with GPT-5.4's quantitative failures even improving slightly (-0.33 pp), and none is significant under a matched-pair McNemar test. It is instead concentrated in the tool-count-mismatch checker ($+10.6$ pp CLAUDE-4.7-OPUS, $+5.3$ pp GPT-5.4, $+5.7$ pp KIMI-2.6) and the textual-reply checker ($+6.2$, $+3.4$, $+3.5$ pp), which together account for essentially all of the $+17.0$/$+9.3$/$+11.1$ pp rise in the overall target-language failure rate. The cross-lingual gap is thus a tool-orchestration and final-response-quality effect, not a numeric- or categorical-reasoning effect, consistent with the ambiguity over-commitment of Section 4.2 (tool-count) and the reply-quality degradation discussed in Section 5 (textual).

G ERROR TAXONOMY

To attribute each multilingual regression to a root cause, we apply a five-category verdict taxonomy, summarised below. Categories are checked top-down; the first that fits is assigned.

1. **Infrastructure (I1–I2)** — The rollout crashed or no judgement was persisted; no gradable attempt exists.
2. **Translation defect (T1–T4)** — The MT pipeline corrupted the task inputs (task description, universe app state field, or oracle event) such that the target-language rollout was set up to fail. Sub-codes cover oracle-event mistranslation (T1), universe app-state corruption (T2), task-description ambiguity or meaning change (T3), and cross-reference/transliteration breakage (T4).[8]
3. **Verifier artefact (K1–K3)** — The agent's action was effectively correct but the verifier rejected it—typically a translation-sensitive *judge* (LLMaaJ) rejection (K1), a locale/format mismatch on a deterministic checker (K2), or list/ordering strictness (K3).
4. **Agent failure (A1–A7)** — The agent genuinely erred on a fair, correctly translated task. It includes: wrong content/args (A1), missing required action (A2), extra/spurious action (A3), wrong tool/target (A4), wrong final answer (A5), refusal/no answer (A6), non-convergence (A7).

[8] In practice T4 is near-empty as a distinct *construction-level* defect: a dedicated non-Latin-script audit finds no transliteration errors introduced by benchmark construction (Appendix J), and the few cross-reference cases we observe share the T1 mechanism—an oracle-compared literal left translated rather than passed through unchanged.

5. **Inconclusive (E1)** — Multiple plausible causes coexist or the deciding evidence (e.g. the judge rationale) is unavailable.

Table 15: Cross-lingual regression cases with per-language failure analysis (CLAUDE-4.7-OPUS, pass@3). Bold marks regressions relative to English.

#	Cap.	English task	eng	spa	cmn	jpn	ind	Code	Per-language failure
1	search	*"events whose **titles** begin with 'Research' "*	3/3	**0/3**	**0/3**	**0/3**	3/3	T3+T2	**spa/cmn/jpn**: "titles" rendered as job/position (*cargos*/职位/役職); agent filters by job title → empty set. **ind**: same T3 (*gelar≠judul*) but agent reads event-title sense; passes via reinterpretation despite T2 word-order break on prefix match.
2	search	*"cities I booked a cab **to** this month"*	1/3	**1/3**	**0/3**	**0/3**	**0/3**	T3+I1	**cmn/jpn/ind**: directional "to" dropped outright; includes out-of-scope pickup city. **spa**: "a las que ... reservar" retains surface form but *reservar* blocks motion reading → agent flags ambiguity, still fails. **eng**: 1 pass + 2 infra aborts (1/1 on completed).
3	exec.	*"Save 13 apartments, average price, email broker"*	3/3	**1/3**	3/3	**1/3**	2/3	K1+A3+A2	**jpn**: judge rejects omitted currency unit in 2/3 rollouts (K1, stochastic; 3rd rollout omits same unit and passes). **ind**: correct end state, but the write-action-count checker reports 14 instead of 13 after an idempotent duplicate (A3). **spa**: agent skips save step entirely despite mentioning it in reasoning (A2, genuine).
4	adapt.	*"Email updated guest list"*	3/3	2/3	3/3	2/3	2/3	K1+A3	**spa**: judge hallucinates misspelling "Dupon" never produced; agent output correct "Dupont" (K1, fabricated rejection). **jpn/ind**: self-correction (omit → redo) reaches correct state, but the write-action-count checker rejects the superseded write (A3).
5	ambig.	*"Save condos; email contacts who wrote 'Dobry den' in August"*	2/3	**0/3**	**0/3**	1/3	1/3	A3+T2	**All langs**: ambiguity correctly handled (agents ask for clarification). **spa/jpn/ind (+eng 1/3)**: write-action-count checker reports 6 instead of 1 because of redundant saves (A3); not cross-lingual. **cmn**: condo property-type enum inconsistently labelled (公寓 vs 共管公寓); filter unreachable (T2) + infra abort.
6	ambig.	*"Delete **my contact** from the US"* (2 US contacts)	3/3	2/3	**0/3**	**0/3**	**0/3**	T3	**eng/spa**: singular+article preserves one-vs-many cue; agent detects mismatch, *asks* (deletes 0). **cmn/jpn/ind**: no article / no obligatory plural → "my contact" reads as generic set; agent perceives no conflict, **deletes both** (irreversible over-action). **spa** (1 failure): agent retrieval miss (finds only 1 contact), not MT.

H REASONING-EFFORT LADDER

Given budget constraints, we run proprietary models at the provider default reasoning effort. In the age of test-time scaling (Snell et al., 2025), it's only fair to ask whether the cross-lingual gap we observe (and report in Section 4) is an artefact of an insufficient *reasoning budget* rather than a

genuine capability deficit. We test this directly for CLAUDE-4.7-OPUS, the strongest proprietary agent, sweeping two of the highest *extended thinking*[9] settings (`high` and `xhigh`) over three target languages, `ita`, `jpn` and `tur`, spanning a Romance-Latin, a Japonic mixed-script and a Turkic agglutinative language and all four capabilities, and study how pass@3 varies with reasoning effort.

Note that exposing a controllable effort setting required re-running CLAUDE-4.7-OPUS under a separate configuration; we therefore compare only *within* the ladder, reporting target-minus-English marginal gaps at a fixed effort. Table 16 reports these gaps at both tested settings. The regression does not close: seven of the twelve target-language $\times$ capability cells retain a significant negative gap, the mean gap over all twelve cells is -7.8 pp in pass@3, and the largest single cell—execution on Japanese, -18.9 pp with a 95% confidence interval of $[-26.9, -10.8]$—is essentially unchanged from its headline value at the very setting the objection predicts should close it (Figure 6). Raising the budget from `high` to `xhigh` narrows the mean marginal gap only marginally (from -9.2 to -7.8 pp) and leaves the sign pattern unchanged, with seven significant negative cells at either setting.

The residual gap is, however, capability-dependent. Execution, the largest gap, is the least responsive to effort (mean -12.4 pp at `xhigh`), while adaptability, whose headline gap is already small, is neutralised (mean $+0.6$ pp, with Turkish even overtaking English by $+4.4$ pp); Japanese is the single cell where raising the budget turns a significant gap non-significant, from -10.7 pp at `high` to -1.9 pp at `xhigh`. Search and ambiguity sit in between (both mean -9.6). We therefore read reasoning effort as a second-order modulator of the cross-lingual gap rather than its cause: at the effort levels we can measure, the majority of the multilingual regression persists, and it is concentrated on the same execution and ambiguity axes that we identified (Section 4.2, Section 4.2).

		`high`			`xhigh`		
Capability	Lang.	pass@3 (%)	Gap (pp)	95% CI	pass@3 (%)	Gap (pp)	95% CI
execution	`eng`	91.8	–		92.5	–	
	`ita`	80.5	**−11.3**	[−18.9, −3.7]	81.1	**−11.3**	[−18.8, −3.9]
	`jpn`	72.3	**−19.5**	[−27.6, −11.2]	73.6	**−18.9**	[−26.9, −10.8]
	`tur`	83.6	**−8.2**	[−15.5, −0.9]	85.5	−6.9	[−14.0, +0.0]
search	`eng`	92.5	–		97.5	–	
	`ita`	85.6	−6.9	[−13.9, +0.0]	88.8	**−8.8**	[−14.8, −3.3]
	`jpn`	81.9	**−10.6**	[−18.0, −3.3]	87.5	**−10.0**	[−16.2, −4.3]
	`tur`	80.6	**−11.9**	[−19.4, −4.4]	87.5	**−10.0**	[−16.2, −4.3]
adaptability	`eng`	76.7	–		75.5	–	
	`ita`	71.1	−5.7	[−15.2, +4.0]	74.8	−0.6	[−10.1, +8.8]
	`jpn`	66.0	**−10.7**	[−20.4, −0.8]	73.6	−1.9	[−11.4, +7.7]
	`tur`	76.7	0.0	[−9.3, +9.3]	79.9	+4.4	[−4.8, +13.5]
ambiguity	`eng`	75.5	–		78.0	–	
	`ita`	66.0	−9.4	[−19.2, +0.6]	66.7	**−11.3**	[−20.9, −1.5]
	`jpn`	61.6	**−13.8**	[−23.7, −3.6]	63.9	**−14.1**	[−23.7, −4.1]
	`tur`	72.8	−2.7	[−12.3, +6.9]	74.7	−3.3	[−12.6, +6.1]

Table 16: CLAUDE-4.7-OPUS reasoning-effort ladder at the two highest tested settings (`high`, `xhigh`): per language, the pass@3 and the target-minus-English marginal gap with its 95% confidence interval, grouped by capability with the English reference (`eng`) as the first row of each block (no gap, "–"). Lang. denotes language. CIs are Wilson–Newcombe intervals on the difference of two independent proportions, and gaps whose CI excludes zero are in **bold** (7/12 at either setting).

I TRAJECTORY TRIAGE PROTOCOL

The automatic estimation of Section 5.1 is produced by a reusable diagnostic procedure that reads the evaluation harness's per-rollout artefacts directly and classifies each failure under the taxonomy of Appendix G. We describe it here in full for reproducibility, together with the infrastructure correction, the stratified reweighting, the whole-dataset extrapolation, and the human validation that back the numbers in the main text.

[9] See Anthropic's extended-thinking documentation.

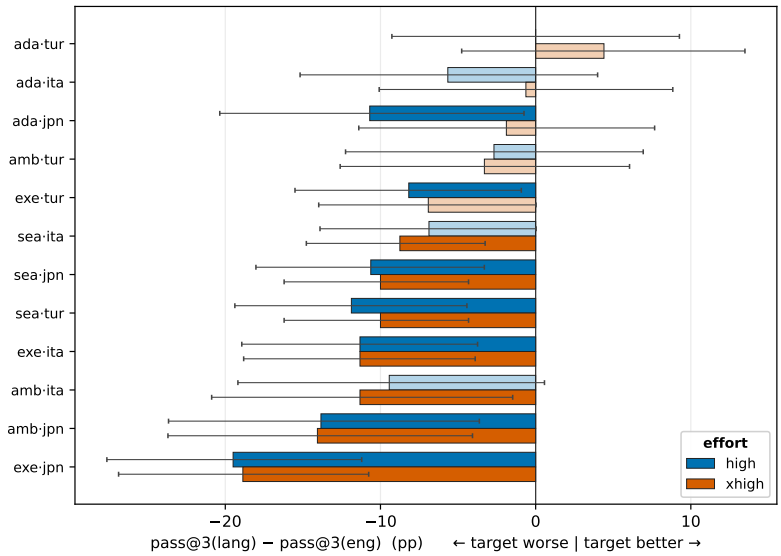

Figure 6: Target-minus-English pass@3 marginal gaps for CLAUDE-4.7-OPUS at both tested effort settings (`high` and `xhigh`), one grouped pair per (capability, language) cell, sorted by the `xhigh` gap. Solid bars mark cells whose 95% Wilson–Newcombe CI on the marginal gap excludes zero; washed bars cross zero.

Stratum	Translation	Verifier	Agent
Deterministic regression (target never solves)	59%	8%	33%
Stochastic regression (target solves some rollouts)	12%	13%	75%

Table 17: Cause composition of genuine target failures by regression stratum. Deterministic regressions are translation-driven; stochastic regressions are agent-driven. Sampling only the former biases the estimate towards translation.

Inputs. For every scenario–language rollout, the harness persists: (i) a *verdict* record with the pass/fail judgement and the failing check(s); (ii) an *environment action log* of the agent's tool calls, from which the graded *write* actions, the task description, and the final answer are recovered; (iii) the raw *agent trajectory* (LLM calls and reasoning); and, in newer dumps, (iv) a *judge log* carrying, per oracle event, the LLMaaJ rationale and the *oracle reference* (the exact expected content). Rollouts are indexed once; the diagnostic agents then address individual rollout directories rather than scanning the corpus.

Phase 1 — worklist construction. From the index we pair each target-language rollout with the English rollout(s) of the same scenario. In *regression mode* we retain scenarios where English passes and the target does not; in *failures mode* (no baseline) we retain failing rollouts directly. Each work item records the rollout directories to compare and a pre-computed failure family (check type, tool-count mismatch, loop/timeout, termination), which supplies a coarse prior before any trajectory is read.

Phase 2 — per-unit diagnosis. One agent processes one scenario (regression mode) or one failing rollout (the finer, stratified mode used for the gap decomposition). It applies the taxonomy *top-down*: infrastructure first (a rollout terminated before a gradable attempt is never blamed on the agent), then translation defect, verifier artefact, and agent failure, with inconclusive reserved for genuinely under-determined cases. The passing English rollout is the baseline the dataset does not otherwise provide: if the target agent took the same correct action yet the verifier rejected it, the verdict is a verifier artefact; if the translated inputs (task description/universe/oracle event) changed the correct action, it is a translation defect; if the inputs are faithful and the agent still erred, it is an agent failure. Following the standard priority rule, a translation defect that also induces a downstream agent error is

Fault side	Share of gap	95% CI
Agent failure	**55.4%**	[50.3, 60.7]
Translation defect	34.5%	[29.5, 39.1]
Verifier artefact	10.1%	[7.1, 13.5]

Table 18: Reweighted, infra-clean decomposition of the cross-lingual gap (CLAUDE-4.7-OPUS, ten target languages, all capabilities). Infrastructure failures (8.5% of rollouts) are excluded; CIs are from a scenario-clustered bootstrap.

attributed to the translation. Judge rejections—otherwise ambiguous between a verifier artefact (K1) and a genuine wrong answer (A5)—are decided by comparing the agent's produced value against the persisted oracle reference and judge rationale, never from which check fired alone.

Phase 3 — compilation and estimation. Per-unit verdicts are aggregated into the fault-side split, the category×capability and per-language breakdowns, and a sub-code distribution. Two corrections make the aggregate gap-representative. First, *infrastructure removal* (below). Second, *stratified reweighting*: writing $R_{\ell,s}$ for the true number of infra-clean failing rollouts of language ℓ in determinism stratum $s \in \{\mathrm{det}, \mathrm{stoch}\}$ (obtained exactly from the rollout index) and $\hat{p}_{\ell,s}(c)$ for the sampled share of cause c in that cell, the per-language composition is

$$\phi_\ell(c) = \frac{\sum_s R_{\ell,s}\, \hat{p}_{\ell,s}(c)}{\sum_s R_{\ell,s}}, \qquad \phi(c) = \frac{\sum_{\ell,s} R_{\ell,s}\, \hat{p}_{\ell,s}(c)}{\sum_{\ell,s} R_{\ell,s}},$$

with $\phi(c)$ the pooled estimate reported in Table 18. Uncertainty is estimated by resampling scenarios (not rollouts) with replacement within each (ℓ, s) cell—a cluster bootstrap, $B = 2{,}000$—which propagates both sampling variance and the correlation among rollouts of the same scenario.

Infrastructure removal. A rollout is labelled *infrastructure* if it carries the harness termination sentinel, a null verdict, or appears in the list of batch-failed (language, capability, rollout) cells arising from per-shard cold starts. Such rollouts persist a non-null but non-gradable outcome, so a naive reading would grade them as agent errors; they account for 8.5% of all rollouts and are excluded from every attribution rate. The batch failures are strongly cell-localised (e.g. a single cold-started shard can void a majority of one language–capability–rollout cell), so they distort per-language scores unevenly if not removed; correcting them was the largest single revision to earlier estimates.

Sampling. For the gap decomposition we triage a language- and capability-balanced sample of infra-clean regressions in each determinism stratum (≈ 4 deterministic and 8 stochastic scenarios per language×capability across the ten target languages), diagnosed at the granularity of individual failing rollouts. For the whole-dataset audit (below) we additionally sample $\approx$11–12 scenarios per language from each of the both-fail and partial→fail cells. All samples use a fixed seed; per-language stratum sizes $R_{\ell,s}$ are taken from the full index, not the sample.

Whole-dataset extrapolation. The triage above conditions on English passing and is by construction blind to scenarios that fail in *both* languages—precisely where a translation defect could deflate a target score without leaving a visible regression. To bound contamination over the entire benchmark we cross-tabulate all 6,056 English×target scenario–language pairs by outcome (Table 19) and focus on the most consequential MT failure mode for cross-lingual scoring: a translation defect that renders the target *unsolvable* (the target never passes). Any pair whose target succeeds on ≥ 1 rollout gives direct evidence against this severe failure mode (74.5% of pairs). The remaining "target-never-solves" pairs partition into three cells—English-solves (pass→fail), English-partial (partial→fail), and English-fails (fail→fail)—each of which we triage; the whole-dataset unsolvable-due-to-translation rate is their size-weighted average. The translation-defect rate falls monotonically as English competence drops (59%, 16%, 10% respectively), yielding the 6.4% bound of Section 5.1 and showing the both-fail blind spot to be cleaner than the visible gap.

Auditing the both-fail cell. Because both-fail scenarios have no passing baseline in either language, we do not judge the agent there. Instead we run an *input-faithfulness* audit: a translation defect is

	Target			
English	pass	partial	fail	**row**
pass	2199	981	435	3615 (59.7%)
partial	400	611	382	1393 (23.0%)
fail	49	271	728	1048 (17.3%)

Table 19: Outcome contingency over all 6,056 English×target scenario–language pairs (CLAUDE-4.7-OPUS, infra-clean, ≥ 2 valid rollouts per side). Pairs whose target passes on ≥ 1 rollout (74.5%) provide direct evidence against MT-induced unsolvability. The three "target never passes" cells—pass→fail (clean regression, 59% translation), partial→fail (16%), and fail→fail (the both-fail blind spot, 10%)—are each triaged; their size-weighted translation rate is the 6.4% whole-dataset bound.

a property of the inputs, so we compare the target task description, universe, and oracle reference against their English source and flag only discrepancies that would change the correct action. This detects translation contamination independently of any pass/fail signal. The audit finds a translation-defect rate of 9.9% (95% CI 5.6–16.9) in the both-fail cell, clustered on a small number of oracle- and keyword-mistranslation bugs (a single scenario accounted for defects in five languages).

Per-language decomposition. Table 20 reports the reweighted composition and the resulting agent-driven share of the conditional gap by language. The agent share is largest for Turkish and Japanese and smallest for Portuguese- and Spanish-heavy comparisons; the translation share is largest for Hindi and Chinese. Per-language intervals are wide (each rests on ≈ 15 sampled scenarios), so they are indicative; the pooled estimate (Table 18) is the reliable quantity.

Language	**Translation**	**Verifier**	**Agent**	**Agent 95% CI**	**Agent gap (pp)**
tur	16%	6%	77%	[62, 90]	20.9
jpn	30%	13%	57%	[40, 74]	17.8
por	26%	6%	68%	[58, 78]	15.2
hin	48%	3%	49%	[31, 67]	14.1
fra	35%	6%	60%	[45, 75]	12.6
cmn	45%	9%	46%	[29, 64]	12.5
deu	43%	8%	49%	[35, 62]	11.0
spa	24%	12%	63%	[49, 78]	10.9
ita	37%	23%	40%	[25, 58]	9.7
ind	38%	15%	47%	[32, 63]	8.9

Table 20: Per-language reweighted gap composition and agent-driven gap (CLAUDE-4.7-OPUS). "Agent gap" is the conditional target failure rate on English-solved scenarios times the agent share. Per-language CIs are wide (≈ 15 scenarios each); the pooled estimate is authoritative.

Reliability and validation. Each verdict carries a self-reported confidence (76% high on the primary regression pass). As an external check, expert linguists independently re-analysed a subset of scenarios (Section 5.2); their attributions on deterministic regressions agree with the automatic labels and with the translation-dominated composition of that stratum, and the deep-dive cases in Table 15 were drawn from this validation. The main residual uncertainty is that the automatic pass tends to *under*-attribute to the verifier (borderline judge cases are conservatively labelled translation defects), so the translation share in Table 18 is a mild upper bound and the verifier-artefact share a lower bound; the agent-driven share, and the whole-dataset contamination bound, are unaffected.

J NON-LATIN SCRIPT ERRORS

We investigate why performance on languages with non-Latin scripts was the lowest, despite these being some of the highest-resource languages (notably cmn). This gap for languages using non-Latin scripts is consistent across all models. Our initial hypothesis was that translation issues were the cause, but in this audit *we find that the script-related failures are not driven by benchmark-construction translation or transliteration defects; the audited scenarios remain solvable.*

We use CLAUDE-5-SONNET, a diagnostic model not included in the evaluated leaderboard, to identify and categorise the agent failures of CLAUDE-4.7-OPUS in cmn, hin, and jpn. Across 939 scenario–language pairs in the three languages, 88 failures were script-related (9.4%). Of these, we categorise 35 (3.7% of all pairs) as entirely the agent's fault and 53 (5.6% of all pairs) as the agent's mishandling of script inconsistencies in the environment data. Both groups belong to the *Agent failure* category because the scenarios remain solvable; the second identifies a subtype in which the agent fails to accommodate cross-script variation in the environment. The most common type of error occurs when the agent searches or filters on only one form of an entity (e.g. written in Chinese characters) and fails because the tool expects it in a different form (e.g. Latin characters). In many cases, the agent explicitly noticed both forms in its own reasoning but still chose to match only one. This class of errors is similar to those identified by Bandarkar et al. (2026). While cross-script settings add a layer of difficulty, a highly capable reasoning model is expected to navigate such situations more effectively (e.g. by retrying the search with multiple forms). We conservatively estimate that this agent's lack of care in multi-script scenarios accounts for about a 2 pp drop in pass@3 accuracy, about half of the gap relative to Latin-script languages such as spa and ind. We anticipate that real-world multilingual agentic settings contain as many, if not more, script inconsistencies in tools and data.

K HUMAN VALIDATION OF TRIAGE LABELS

To verify the automatic attribution (Section 5.1) at scale, we had native or fluent speakers re-adjudicate a sample of triaged failures through a purpose-built review interface, independently of the automatic protocol.

Protocol. Each item pairs one failing target-language rollout with its automatic verdict. The reviewer sees only the decision-relevant evidence—the English task description, its target-language translation, the oracle event, the LLMaaJ rationale (where applicable), and the agent's final answer—rather than the full trajectory. For every item the reviewer either *confirms* the assigned fault side, *rejects* it and names the side they believe is correct, or marks it *undecidable* when the evidence is insufficient. The three sides mirror the taxonomy of Appendix G: translation defect (T), verifier artefact (K), and agent failure (A); infrastructure and inconclusive items are not sampled. Items are drawn from both regression strata (deterministic regressions, internally labelled gap, and stochastic regressions, internally labelled partfail) and balanced across languages and capabilities.

Coverage. Five human reviewers contributed **149** items in seven languages (spa, ita, hin, jpn, ind, por, cmn) and all four capabilities. Excluding the 9 undecidable items, **128 of 140** adjudications confirmed the automatic label—**91.4%** agreement (95% Wilson CI 85.6–95.0). Agreement is comparable across fault sides (Table 21) and strata (gap 92.9%, partfail 88.1%). Per-language agreement (Table 22) is uniformly high except for Japanese (68%), where reviewers reattributed several agent-failure labels to translation defects—a small, script-specific pocket we flag for follow-up rather than a systematic bias.

Fault side	Adjudicated	Confirmed	Agreement
Translation defect (T)	22	19	86.4%
Verifier artefact (K)	27	25	92.6%
Agent failure (A)	91	84	92.3%
All	**140**	**128**	**91.4%**

Table 21: Human re-adjudication of automatic triage labels by fault side (undecidable items excluded). Agreement = confirmed / adjudicated.

Language	Adjudicated	Confirmed	Agreement
Spanish	29	27	93.1%
Italian	26	25	96.2%
Hindi	23	23	100.0%
Japanese	22	15	68.2%
Indonesian	22	22	100.0%
Portuguese	10	9	90.0%
Mandarin	8	7	87.5%
All	**140**	**128**	**91.4%**

Table 22: Human re-adjudication agreement by language (undecidable items excluded).