Actions Speak Louder than Words: Measuring Cross-Lingual Policy Retention in Tool-Using Agents

Sourabrata Mukherjee **Kalika Bali** **Sunayana Sitaram**
Microsoft Research India
t-somukherje@microsoft.com

Abstract

When a tool-using agent is given the same task in a different language, does it still take the same steps? Multilingual evaluation rarely asks, because it compares final answers and throws the intermediate actions away. For an agent those actions are the product: the route fixes cost and latency, decides how the system fails, and is the only part of its behaviour a reviewer can audit, so two language versions that agree on every answer can still differ in price, in failure mode, and in whether a safeguard written against one of them holds for the other. We therefore make the action policy itself the measured object, across 8 models, 6 parallel benchmarks and 41 languages (2.38M agent rollouts). The naive measurement does not work, because five confounds sit between raw trace similarity and any defensible claim, each large enough on its own to change a conclusion: short traces score higher than long ones, empty traces score perfectly, unrelated traces already agree by chance more than half the time, the gap is capped by each model's own reproducibility, and, worst of all, a model asked the same question twice in one language does not answer it the same way, so there is no reference point to measure against. We rebuild the measurement to remove all five, and every correction makes the effect larger rather than smaller. Cross-lingual divergence then proves structural rather than sampling noise: it survives greedy decoding in every cell we test and stays flat as temperature rises, even as models become far less self-consistent. Normalising by each model's own reproducibility, four very different frontier models converge on nearly the same value *under greedy decoding*, each keeping 71–73% of its action policy when the language changes, and which model it is explains only 5.7% of the variance. Below roughly 10B parameters that regularity breaks down, and the apparent ordering among smaller models is largely an artifact of a chance floor that we measure by permutation rather than assume. Asking why, we find that agents route non-English tasks through English, that this pivot is causally load-bearing, with a prediction registered in advance and confirmed across four models, and that models will not abandon it when instructed to. Finally, a single trace-extraction regex, not the model, manufactured an apparent multilingual failure: two worked examples raise one model's measured accuracy twenty-sixfold while its accuracy on readable outputs barely moves. Code, prompts, both benchmark suites and all 2.38M per-rollout traces are released.[1]

1 Introduction

Language models are increasingly deployed as *agents*: they break a task into steps, call tools, and assemble a result (Yao et al., 2023; Schick et al., 2023; Parisi et al., 2022; Qin et al., 2024). Ship one multilingually and the question that matters is not only whether it answers correctly in Hindi, but whether it *does the same thing* in Hindi as in English: whether it retrieves before it computes, whether it translates first, whether it takes three steps or eight.

[1]Code, prompts and results: repository. Data: the adapted and synthetic multilingual agentic tool-use benchmarks, both on Hugging Face. Key terms are defined in Table 5, Appendix A.

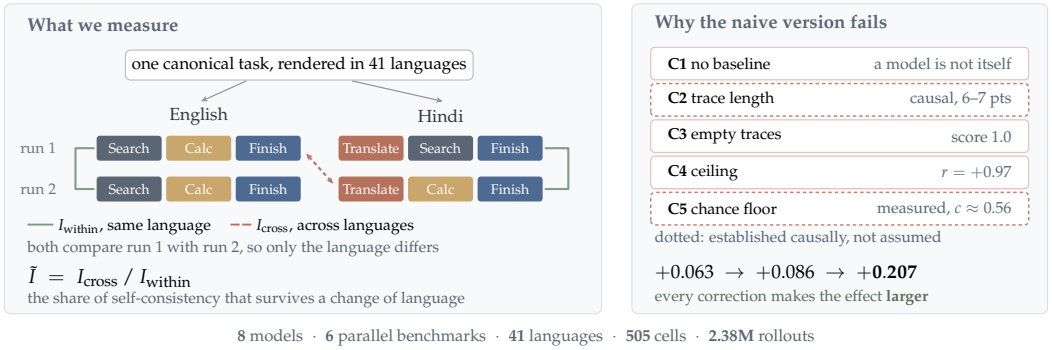

Figure 1: Semantically identical tasks in 41 languages pass through one fixed symbolic tool-use scaffold, and every cell is generated twice so that same-language and cross-language agreement are estimated the same way. The traces are illustrative: the Hindi arm opens with the English pivot, and the two runs of a single language do not match each other either, which is why a same-language baseline is needed before any cross-language number means anything. Five confounds sit between raw trace similarity and a defensible claim, four of them large enough on their own to reverse a sign or a ranking.

That difference is not cosmetic, because for an agent the route is the product. A model that translates a passage before answering spends tokens and latency that the English arm never spends, so the identical request costs more in one language than in another. A different route also fails differently: an error introduced by that translation step exists only on the non-English path and will never surface in an English regression test. And the route is the part of an agent's behaviour that can actually be governed, since tool permissions, rate limits, escalation rules and audit trails are all written against the sequence of actions a system is expected to take, so a policy validated on the English trace does not describe what the system does in Hindi. None of this means a different language should always produce an identical route; sometimes it should not. It means the difference has to be measurable, and today it is not measured at all, because answer-level parity is compatible with any amount of it.

Multilingual evaluation has almost no purchase here. Standard suites compare final outputs (Hu et al., 2020; Ahuja et al., 2023; Liang et al., 2023), and agentic benchmarks are overwhelmingly English (Liu et al., 2024; Mialon et al., 2024), so the intermediate policy is discarded on both sides. We make that policy the measured object and ask: **do multilingual tool-using models execute the same action policy for semantically identical tasks across languages?** Figure 1 summarises how we answer it.

What is hard here is not collecting the traces but interpreting them: rendering a task in many languages, extracting the traces and averaging their similarity produces a number that cannot be trusted, for five separate reasons. Four are metric artifacts: short traces score higher than long ones, empty traces score perfectly, unrelated traces already agree more than half the time by chance, and the gap is bounded by a model's own reproducibility. The fifth is deeper. There is no baseline, because a model asked the same question twice in its own language does not answer the same way, and without that reference point the quantity is not identifiable at all.

Our protocol fixes this. Under a fixed scaffold with a shared five-tool alphabet, a model emits an executed action trace alongside its answer, and the trace is what we compare. We generate every cell *twice* under identical decoding, task set and token budget, varying only the serving seed. Same-language agreement I_{within} pairs the two replicates in one language; cross-language agreement I_{cross} pairs them across two. Both sides are cross-seed, so decoding noise enters identically and language is the only difference. **Normalised policy retention** is $\tilde{I} = I_{\text{cross}}/I_{\text{within}}$, the share of a model's own reproducibility that survives a change of language. Every comparison is length-matched in both directions and accepted only when both agree in sign; pairs are dropped whenever either trace is empty; intervals

are task-level bootstraps. The estimator is set out in full in Appendix B, the five confounds with their measured cost in Appendix E, and the design checks in Appendix H.

Corrected this way, the effect grows rather than shrinks. Every correction we apply makes it larger, never smaller. Under greedy decoding the gap is positive in all 24 cells, with every bootstrap interval on the raw gap excluding zero (Figure 2), and a five-point temperature ladder shows why sampling had been hiding it: cross-lingual agreement barely moves as temperature rises, while self-consistency falls away sharply.

Dividing by that ceiling gives the paper's central result. Under greedy decoding, four models that share almost nothing, an order of magnitude of scale, two architectures and entirely different training mixtures, converge on nearly the same value: each retains **71–73%** of its own action policy when the language changes. The regularity holds per cell rather than only per model, and across all 24 of them model identity explains barely any of the variance (5.7%). Extending the study to three smaller and other-vendor models shows where this regularity ends, and why the ordering among small models cannot be read off the raw metric at all (§4).

Asking why it happens points to English. Agents route non-English tasks through an English pivot: `Translate` is the most-used tool in every adapted benchmark, and reasoning text is ≈99% ASCII even on Devanagari input. Removing the translation tool lowers length-matched agreement in proportion to how much a model uses it, and mandating it helps in a **pre-registered** ordering across four models, predicted in advance from each model's measured head-room. The pivot survives a direct instruction to abandon it, at under 1% compliance. This connects interpretability findings about latent English processing (Wendler et al., 2024; Zhao et al., 2024) to a causal test at the level of executed actions. Separately, a single trace-extraction regex, not the model, manufactured an apparent multilingual failure in GPT-OSS-120B, which we take apart in §6.

The work sits between two literatures that rarely meet. Agentic benchmarks built on ReAct-style scaffolds (Yao et al., 2023; Schick et al., 2023; Liu et al., 2024; Mialon et al., 2024) score *outcomes* and are overwhelmingly English, while cross-lingual suites (Hu et al., 2020; Ahuja et al., 2023; Liang et al., 2023) compare final predictions and multilingual chain-of-thought work (Shi et al., 2023; Muennighoff et al., 2023) compares free-form *text*, which is hard to align. We make the executed trace the measured object, which a shared symbolic action alphabet renders comparable. Closest in spirit, Qi et al. (2023) ask whether models retrieve the same *facts* across languages; we ask it of *procedures*, and add the matched baseline without which the quantity is not identifiable (Appendix C, with an extended comparison in Appendix D).

We contribute a ceiling-corrected estimand for cross-lingual policy retention, and a measurement protocol that prices each of five confounds on its own data, two of them established causally rather than assumed (Appendices B and E). We apply it at a scale that lets the question be answered rather than gestured at: 2.38M rollouts over 505 cells, spanning 8 models, 41 languages, five temperatures and six interventions (§2). We show the retention regularity is real, locate its boundary, and demonstrate that the apparent ordering among models below that boundary is a metric artifact (§4). And we establish the mechanism causally, with the key prediction registered in advance (§5).

2 Experimental Setup

Comparing policies across languages requires $x^{(\ell_i)}(z)$ and $x^{(\ell_j)}(z)$ to be the *same* task, a stronger condition than it looks and one that eliminates most multilingual corpora. En–X bitext pairs each language with English but not with the others, so row i of the Hindi arm and row i of the Bengali arm are unrelated sentences. Of the corpora we screened, five fail this test outright, and a sixth stores its per-language configurations in different row orders, so that an index join matches only half the items. The suite we keep is six benchmarks on verified alignment keys (FLORES-200 (NLLB Team et al., 2022), XQuAD (Artetxe et al., 2020), XNLI (Conneau et al., 2018), Belebele (Bandarkar et al., 2024), XCOPA (Ponti et al., 2020) and a purpose-built synthetic benchmark): **5,776 tasks**, **41 languages** across 17 families and

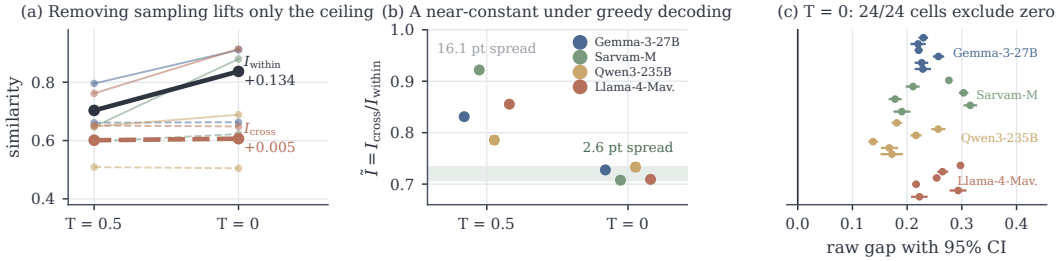

Figure 2: The central result. (a) Greedy decoding raises I_{within} by 0.134 but moves I_{cross} by 0.005; thin lines are models, thick lines the pooled mean. (b) Dividing by each model's ceiling collapses a 13.6-point spread at T=0.5 into a 2.6-point band at T=0, with task-level bootstrap 95% intervals. (c) All 24 cells at T=0 are positive, with every interval on the raw gap excluding zero.

Estimator	I_{within}	I_{cross}	Raw gap	**Length-matched**	Cells > 0
Unmatched token budget	0.6634	0.5370	+0.1264	+0.0625	14/15
Matched budget, single-replicate pairing	0.7103	0.6049	+0.1054	+0.0878	16/16
Matched replicates, lenient exclusion	0.6821	0.5711	+0.1110	+0.0946	18/18
Matched replicates, T=0.5	0.7033	0.6011	+0.1023	**+0.0861**	**24/24**
Matched replicates, T=0	**0.8371**	0.6064	+0.2307	**+0.2074**	**24/24**

Table 1: Every correction makes the effect larger. Pooled over the four protocol-adherent models, with GPT-OSS excluded and analysed in §6. The uncorrected baseline is shaded orange and the two headline estimates blue. The last two rows cover identical models and tasks and differ only in temperature; in both, every 95% bootstrap interval on the raw gap excludes zero.

16 scripts, and 809 language pairs. We call the five repurposed public benchmarks *adapted* and the purpose-built suite *synthetic*. Per-benchmark statistics, the language inventory and every alignment key are in Appendix F.

Five instruction-tuned models form the main suite, spanning scale, architecture and language specialisation: **Gemma-3-27B** (Gemma Team, 2025) and **Sarvam-M** (24B, Indic-specialised) are dense; **Qwen3-235B-A22B** (Yang et al., 2025) and **Llama-4-Maverick** (17B active, 128 experts) are mixture-of-experts; **GPT-OSS-120B** (OpenAI, 2025) is the fifth. To locate the boundary of the regularity in §4 we add **Gemma-3-4B** (same family and recipe, 6.75× smaller, isolating scale), **Qwen3-8B** (same family, different training mix) and **Aya-Expanse-8B** (Dang et al., 2024) (a fourth vendor, explicitly multilingual post-training), giving **eight models in total**. Every task–language instance uses an identical scaffold requiring `Thought: / Action: ToolName(args)` steps ending in `Finish`, with traces extracted by a single regex. Tools are *symbolic*: calls are parsed and compared but never executed, so we study induced tool-use policy rather than grounded execution. Unless stated otherwise decoding is T=0.5, `max_tokens`=4096 and an iteration cap of 10; serving is vLLM (Kwon et al., 2023) on 8×B200 nodes.

The estimand needs replicates, the confounds need ablations, and the boundary needs models outside the frontier, so the study totals **2,382,875 rollouts across 505 cells and 8 models** (Table 8). Three checks precede every claim: language routing is verified from each run's own prompt log, where prompt sharing between arms is **0.0%**; truncation at 4096 tokens is at most 0.63% for every model; and coverage is exact, with all 44 arms of the *audited* campaign holding exactly the planned cells and rollouts (Appendix V).

3 Cross-Lingual Policy Divergence

Table 1 tracks the estimate as each correction is applied. An unmatched baseline, the value an unwary reuse of existing runs would produce, reports +0.0625; matching the budget raises it to +0.0878, and strict empty exclusion with a fourth model added gives +**0.0861**. **The matched estimate is 38% larger than the unmatched one.** The natural worry is that a confound might be *creating* the effect; here every confound was *suppressing* it.

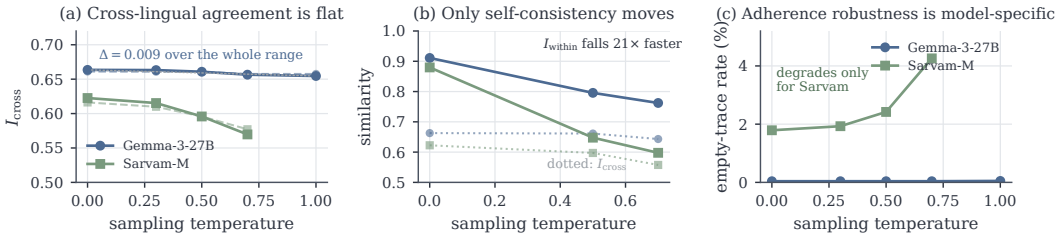

Figure 3: The temperature ladder separates language from sampling. (a) Cross-lingual agreement is flat from greedy to T=1.0; solid is raw, dashed is length-matched. (b) Self-consistency (solid) falls steeply while I_{cross} (dotted) barely moves, so the narrowing gap at high temperature is a ceiling effect rather than convergence. (c) Adherence under temperature is model-specific and matters for deployment: Gemma holds a 0.04% empty-trace rate throughout, whereas Sarvam degrades 2.4× (Appendix M).

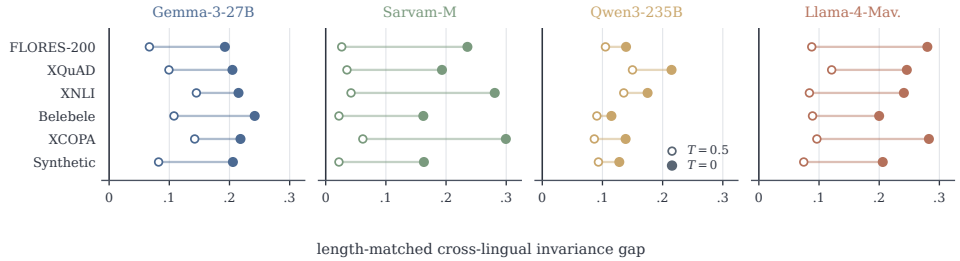

Figure 4: Each rule is one (model, benchmark) cell, running from its sampling-inclusive gap (open) to its greedy gap (filled). All 48 are positive and all 24 rise under greedy decoding, so no single benchmark or model carries the effect, and at T=0 every bootstrap interval on the raw gap excludes zero. Numeric values and intervals are in Appendix I.

If cross-lingual differences were decoding noise correlated with language, removing the noise should shrink the gap. It does the opposite: on the same four models and tasks the length-matched gap rises to $+\mathbf{0.2074}$, again positive in 24/24 cells, with every bootstrap interval on the raw gap excluding zero. The decomposition (Figure 2a) is unambiguous, since removing sampling lifts same-language self-consistency sharply and leaves cross-language agreement almost untouched. **Sampling noise was masking the language effect, not producing it**, and $+0.0861$ is a lower bound on a sampling-free effect of $+0.2074$.

Two points do not make a curve, so we ran $T \in \{0, 0.3, 0.5, 0.7, 1.0\}$ (Figure 3). Cross-lingual agreement is essentially flat across the entire sampling range, while over the sub-range where both quantities are measurable, self-consistency moves twenty-one times further. This licenses a sharper statement than "the gap doubles at T=0", which is partly definitional since the gap tends to $1 - I_{\text{cross}}$ as decoding becomes deterministic: **cross-lingual policy divergence is invariant to sampling temperature**. It is a property of how the model maps language onto actions rather than of the sampler, and the apparent temperature-dependence of the *gap* is the ceiling moving, the failure mode C4 predicts. The T=0.7 arm, paired against a separate five-replicate experiment, also reproduces $\tilde{I}$ to within 0.003 by an independent route.

The effect is uniform across cells: all 48 length-matched gaps are positive (Figure 4), including on the synthetic benchmark, whose traces are by far the longest (Appendix I). Per-model summaries (Table 2) show the four models differ substantially in *absolute* gap, precisely the difference §4 shows is not what it appears to be.

4 Policy Retention and Its Boundary

A model that cannot reproduce itself has no room to display a cross-lingual gap. Across the four adherent models, the correlation between a model's own self-consistency and its measured gap is $r = +0.43$ at T=0.5 and $\mathbf{r} = +\mathbf{0.97}$ at T=0: at greedy decoding, essentially

	$T = 0.5$ (sampling-inclusive)				$T = 0$ (greedy)			
Model	I_w	I_c	Len-match	$\tilde{I}$	I_w	I_c	Len-match	$\tilde{I}$
Gemma-3-27B	0.7971	0.6734	+0.1072	0.8311	0.9060	0.6754	+0.2131	**0.7276**
Sarvam-M	0.6304	0.5878	+0.0349	0.9219	0.8632	0.6175	+0.2224	**0.7076**
Qwen3-235B	0.6375	0.4977	+0.1101	0.7855	0.6787	0.4902	+0.1517	**0.7331**
Llama-4-Maverick	0.7484	0.6454	+0.0922	0.8553	0.9005	0.6424	+0.2425	**0.7092**
spread			*0.075*	*0.136*			*0.091*	***0.026***

Table 2: Per-model results at both temperatures. I_w and I_c are macro-averaged over the six benchmarks whereas $\tilde{I}$ is computed on pooled pairs, so dividing the two printed columns reproduces $\tilde{I}$ only to within 0.02; we quote the pooled figure throughout and reconcile both aggregations in Appendix I. The absolute gap spreads models widely; the normalised estimand collapses them at T=0.

Grouping	Mean $\tilde{I}$	Range	Width
by benchmark			
Belebele	0.776	0.730–0.815	0.085
XCOPA	0.720	0.647–0.776	0.129
XQuAD	0.719	0.614–0.776	0.161
FLORES-200	0.713	0.675–0.756	0.081
XNLI	0.712	0.659–0.771	0.112
Synthetic	0.691	0.676–0.705	0.028
by model			
Gemma-3-27B	0.742	0.676–0.772	0.095
Qwen3-235B	0.720	0.614–0.794	0.180
Llama-4-Maverick	0.713	0.682–0.767	0.084
Sarvam-M	0.713	0.647–0.815	0.168
All 24 cells	**0.722**	0.614–0.815	SD **0.052**

	T=0	T=0.5
variance in $\tilde{I}$ explained by (η^2)		
Model identity	5.7%	74.8%
Benchmark	26.9%	6.9%
Residual	67.4%	18.3%
spread across the four models		
Band width in $\tilde{I}$ (pts)	2.6	13.6
Relative spread	3.5%	16.1%
rank by uncorrected absolute gap		
Qwen3-235B	4	1
Gemma-3-27B	3	2
Llama-4-Maverick	1	3
Sarvam-M	2	4

Table 3: Policy retention per cell at T=0 over $n = 24$ cells, and what moves it. Left: retention by benchmark and by model; the ranges overlap heavily, which is the per-cell form of the claim. Right: one-way η^2, the width of the across-model band, and the four models ranked by their uncorrected absolute gap. Under greedy decoding model identity explains almost nothing and the benchmark dominates; at T=0.5 that reverses and the ranking inverts with it ($\rho = -0.80$), on identical models and tasks. Per-cell predictors are in Appendix J.

all between-model variation in the headline gap is explained by the models' own reproducibility rather than by their cross-lingual behaviour. Any ranking built on the absolute gap is therefore a ranking of determinism wearing a multilingual label, and the consequence is concrete. Ordering models by the T=0.5 gap and by the T=0 gap gives rank correlation $\rho = -0.80$ (Table 3): two temperatures, the same models, near-opposite orderings. **No per-model ranking derived from an uncorrected cross-lingual gap should be published.**

Dividing by the ceiling removes the contamination. Under $\tilde{I} = I_{\text{cross}} / I_{\text{within}}$, bootstrapped directly at the task level, the four models land within **2.6 percentage points** of each other at T=0, every interval narrower than the band itself (Table 2, Figure 2b). A 27B dense model, a 24B Indic-specialised dense model, a 235B mixture-of-experts and a 128-expert mixture-of-experts, spanning an order of magnitude in scale, two architectures and entirely different training mixtures, **all retain 71–73% of their own action-policy self-consistency when the language changes**. The relative spread falls from 16.1% at T=0.5 to 3.5% at T=0.

The obvious objection is $n = 4$, and the data already answers it. Because $\tilde{I}$ is defined per cell there are 24 of them, and across all 24 it has an SD of just 0.05. Decomposing that variance is the informative step. **Model identity explains 5.7% of it; the benchmark explains 26.9%.** That figure is small in absolute terms, and smaller still against the same estimand on the same cells under sampling, where model identity explains 74.8%, so removing the sampler collapses the model term thirteenfold. The claim is therefore stronger than "four models agree": **policy retention is close to a model-independent quantity**, and what looks like a per-model multilingual characteristic at T=0.5 is largely each model's own sampling noise. The regularity is specific to greedy decoding, so any use of the 71–73% figure must state the temperature. Correcting for the measured chance floor lowers the level roughly

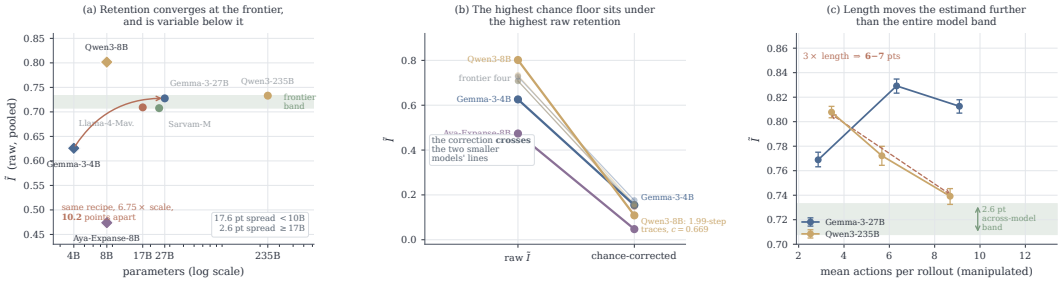

Figure 5: The regime, its boundary, and why the boundary is a measurement result. (a) $\tilde{I}$ against parameter count, with the shaded band marking the frontier range; the arrow marks the within-family comparison of same recipe, 6.75× scale, ten points apart. (b) Raw against chance-corrected $\tilde{I}$, where the correction crosses the two smaller models' lines. (c) The three-level length manipulation moves $\tilde{I}$ further than the entire across-model band.

fivefold and preserves the *absolute* band, though not the relative one, since a fixed spread on a five-times-smaller base is a five-times-larger relative spread. We report both, and say which the claim rests on (Appendices I and Q).

One model is anomalously irreproducible under greedy decoding, and the natural explanation, expert routing in mixture-of-experts serving, is refuted by the fourth model, which is also an MoE and among the most reproducible in the study. Trace length is the strongest per-cell predictor, and §5 establishes that it is causal (Appendix J.1).

Four models is a striking regularity but a bounded claim, so we extended the study to three systems chosen so that a break would tell us something. One shares a family and training recipe with a model already in the set but is nearly seven times smaller, which isolates scale. One shares a family but differs in training mixture. One comes from a fourth vendor and was explicitly post-trained for multilinguality. **At least two of the three fall outside the band, on every way we tried to compute it** (Figure 5a).

It is not a scaling law. The sharpest line available is the within-family comparison: same recipe, 6.75× the parameters, and the two models sit ten points apart. But scale does not order the small models either. Corrected for chance, the 4B model lands *inside* the frontier band while the 8B model falls below it, which is the opposite of what a scaling account predicts. What replaces the constant is a regime: retention converges at the frontier and is variable below it, an eightfold difference in SD, which we offer as a hypothesis at $n = 2$.

This is where measuring the chance floor pays. One 8B model emits the shortest traces in the study and therefore carries the highest floor of any model measured: two of its traces answering *different* tasks already score 0.669, and 0.947 on one benchmark where its cross-language agreement is 0.9497. Its frontier-beating raw retention is largely that floor. Corrected, it falls out of the band while the 4B model moves in, so the correction does not merely rescale the numbers, it reverses which of the two smaller models looks better (Figure 5b), and no statement about small-model retention is interpretable on the raw scale.

The obvious alternative is that the frontier four merely occupy a narrow band of trace lengths. One model refutes this structurally: its mean trace length sits *inside* the frontier range and it still misses by eight points. Across the six protocol-adherent models length explains under a third of the variance in retention, so it is a moderate correlate rather than the explanation. We report the fourth-vendor model as an *adherence* failure rather than a retention datapoint, since it emits no parseable trace on nearly a third of its rollouts, which leaves the fourth-vendor question open. It also produces an observation worth stating plainly: the two systems in this study specialised for multilinguality are the two that do worst at multilingual agentic behaviour, one on protocol adherence (Appendix R) and one on English advantage (Appendix P).

		`Translate` *removed*				`Translate` *mandated*				
Model	Head-room	Raw Δ	A	B	Bench	1st act.	Raw Δ	A	B	Bench
Sarvam-M	**57.2**	+0.020	+0.0015	+0.0064	1/6	81.4%	−0.005	+0.0110	+0.0190	**5/6**
Gemma-3-27B	30.4	+0.023	**−0.0432**	**−0.0348**	4/6	99.6%	+0.032	**+0.0635**	**+0.0674**	**5/6**
Llama-4-Maverick	18.7	+0.055	−0.0511	+0.0124	4/6	98.7%	−0.007	+0.0242	+0.0154	3/6
Qwen3-235B	15.7	+0.062	−0.0229	+0.0659	4/6	99.0%	+0.073	−0.0105	−0.0117	2/6

Table 4: The English-pivot ablation on all four models, ordered by head-room, which is 100 minus the baseline share of non-English rollouts opening with `Translate`; "1st act." is that share under the mandate. A and B are the two length-matching directions on I_{cross} and "Bench" counts the benchmarks of six signed as predicted with both agreeing. **Every removal arm looks beneficial raw, yet three of the four are harmful once length is matched** (blue), on the per-benchmark reading that the orange cells make necessary: those two pooled estimates fail the both-directions test, as does Sarvam's partial compliance. Detail in Appendix N.

5 The English Pivot

Two observations point the same way. `Translate` is the *most frequently used tool in every adapted benchmark*, for every compliant model, while the synthetic benchmark, designed around arithmetic composition, correctly inverts to `Calc`. And the models' own reasoning text is ≈**99% ASCII even when the prompt is Devanagari, Tamil or Odia**, although the prompts are verifiably in-language (§2). Agents translate into English, plan in English, and answer: the behavioural counterpart of interpretability results showing that multilingual transformers pivot through English-aligned latent representations (Wendler et al., 2024; Zhao et al., 2024), but unlike hidden-state evidence it can be tested by intervention. The reliance is uneven, with Qwen3 spending **50.5%** of its tool calls on `Translate` against Sarvam's 18.0%, which is what makes the mechanism testable rather than merely observable.

We manipulate the pivot with two matched arms on *four* models, holding seed, temperature, budget and task set fixed and changing only the tool alphabet or an ordering constraint: `Translate` *removed* from the offered tools, and `Translate` *mandated* as the first action on non-English tasks. **Seven of the eight manipulations took cleanly**: removal eliminated the tool completely in all four models, none of which reached for a tool it was not offered (6 calls in ∼500,000), and the mandate took on **98.7–99.6%** of non-English rollouts in three of them. Sarvam-M is the exception at 81.4%, and we flag it wherever its magnitude is quoted. Removal causes *substitution rather than omission*, with a model-specific substitute: Gemma into `Search/Calc`, Qwen3 and Llama-4 into `Summarize`, Sarvam into `Search` (Figure 6b).

The raw numbers invert the conclusion. Read without length control, *every* arm that *deletes* the pivot appears to help, on all four models (Table 4). Removing a tool shortens traces, and shorter traces score higher similarity, so the raw estimate simply reads that compression back. Length-matched in both directions (Figure 6a), removing the English pivot lowers cross-lingual agreement in proportion to how much a model uses it: the effect holds per benchmark for the three models that pivot heavily and not for the one that pivots least. This is a dose–response refinement of the mechanism rather than a counterexample to it, and the substitution pattern points the same way, since withdrawing the tool sends the freed budget to a different English-producing operation (Appendix N).

Our first reconciliation of *why* the mandate helps some models and not others was post-hoc, so we tested it in advance: two further models were chosen on their measured baseline pivot rate, one with more head-room than any model tested and one with almost none, and both predictions were written into the analysis script before the compute was spent. Across four models the mandate's benefit is **monotone in head-room**, at 5/6, 5/6, 3/6 and 2/6 benchmarks (Table 4). Head-room governs *whether* it helps, not how much, since the model that moved furthest gained 4.4× less than the one that moved second-furthest, so we report the direction as tested and offer no general prescription.

If English reasoning were a stylistic default, one scaffold line should change it. We tested two further arms: write every `Thought:` in English, or in the task's language. **The second manipulation was refused.** Told explicitly to reason in the task's language, Gemma writes

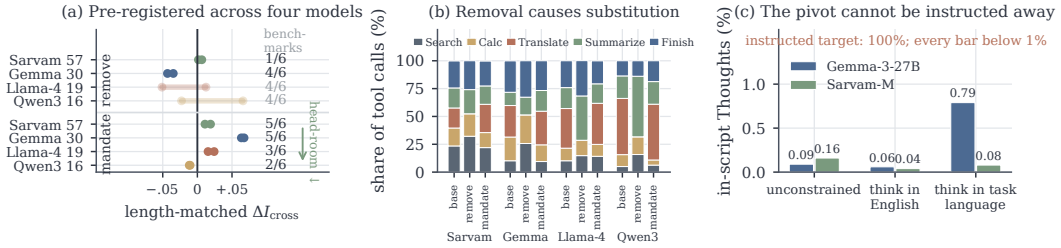

Figure 6: English pivoting, tested causally and pre-registered across four models. (a) Length-matched ΔI_{cross}; two dots per row are the two matching directions, a faded bar means they disagree, the number beside each model is its head-room, and the right column counts benchmarks of six signed as predicted. The mandate's *benefit* is monotone in head-room (5/6, 5/6, 3/6, 2/6); its *magnitude* is not. (b) Removal causes substitution, not omission. (c) Instructing models to reason in the task's language fails, at under 1% compliance against a target of 100%.

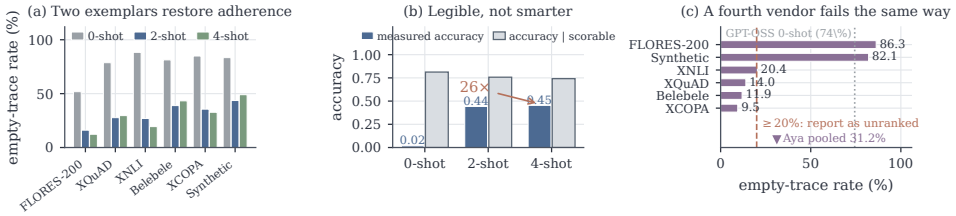

Figure 7: Measurement pathology and its repair. (a) GPT-OSS-120B's empty-trace rate per benchmark under 0/2/4-shot exemplars. (b) Measured accuracy rises 26× while accuracy among scorable rollouts is flat, so the repair is legibility rather than capability. (c) The pathology is not one model's quirk: Aya-Expanse-8B, a fourth vendor, emits no parseable trace on 31.2% of rollouts, and on more than four in five for its two worst benchmarks. Dashed: the 20% unranked threshold; dotted: GPT-OSS's 0-shot rate, for reference.

a non-Latin-script thought on **0.79%** of non-Latin-script rollouts and Sarvam on **0.08%** (Figure 6c). We report this as a *failed manipulation*, not a null: because the instructed condition was never achieved, the experiment cannot test whether reasoning language affects invariance. What it establishes is stronger than the null it replaces: **the English pivot is not a preference that prompting can switch off**, surviving a direct instruction in two models at a refusal rate above 99%. That also explains the asymmetry above, since a scaffold line can only *add* pivoting where it was absent and cannot remove it.

Across the adherent cells, mean trace length alone explains two-thirds of the variance in raw invariance, but that is correlational, so we manipulated it. A three-level intervention holding task, language, seed and budget fixed moves mean trace length roughly threefold, and moves $\tilde{I}$ by **6–7 points with disjoint intervals, further than the entire band across the four frontier models** (Figure 5c). So the normalisation does *not* absorb length, and benchmark-level claims resting on it being length-neutral must be withdrawn; length is not a nuisance to be corrected away but a first-order driver of agentic policy retention, which is why every comparison in this paper is length-matched in both directions. The two models tested disagree on the *sign*, so we make no universal claim about the direction (Appendix J.1).

6 Outcomes and Measurement Validity

A reasonable objection is that traces are means, not ends: if the answers are right, who cares how the agent got there? We scored correctness against gold on the four benchmarks that have it, with 14,000 gold-bearing rollouts per condition. The answer cuts both ways.

The language effect does reach the outcome. Cross-lingual accuracy spread reaches **0.66 within a single model and benchmark**: Qwen3 scores 0.990 on its best XCOPA language

and 0.330 on its worst, and Sarvam spans 0.278 to 0.789 on XQuAD (Figure 9a). Averaged over the four gold benchmarks every model has an English advantage, and the ordering is counter-intuitive: **Sarvam-M, the Indic-specialised model, has the largest** at $+0.155$ (Figure 9b). Specialising a model for a language family does not, on this evidence, close the English gap in *agentic* performance (Appendix P).

Invariance is nonetheless not a proxy for accuracy. Pooled across all cells the correlation looks decisive ($r = +0.897$), and it must not be used: it is a line drawn through two clusters, one of which is GPT-OSS measuring a different variable entirely (Figure 9c). Among protocol-adherent models the association is positive but moderate, and it *reverses* in a quarter of cells. Trace invariance and correctness must be measured separately.

The empty-trace rate that C3 forces us to report separately is itself a finding, and it collapses deterministically for low-resource languages. One model emits no tool call at all on a fifth of the hardest task family, and the failures are neither truncation nor noise: the same task in the same language fails in both independent replicates. They concentrate on low-resource Indic languages, and 22 of 23 are affected. This is distinct from policy divergence, because the model does not act differently; it fails to act at all.

The sharpest case is a regex that made a competent model look broken. GPT-OSS-120B yields no parseable trace on 76.4% of its rollouts and scores under 2% accuracy, yet posts the two *highest* raw invariance scores in the study, because empty pairs score 1.0. The model is not broken; the parser is. It simply answers in prose that names the tool, writing "We will use Translate." instead of emitting the required syntax, and most of its non-empty failures are machine-recoverable. Two worked exemplars, with model, decoding and tasks fixed, raise its measured accuracy **twenty-sixfold** while its accuracy on readable outputs barely moves (Figure 7): the intervention made it legible, not smarter, which raises the question of which of the two numbers was ever measuring capability. Single-pattern extractors are standard in ReAct-style harnesses (Yao et al., 2023; Liu et al., 2024) and parse failure need not be language-uniform, so such an artifact can masquerade as a multilingual finding; we recommend reporting parse-failure rate alongside every headline number and treating any model above $\sim$20% as unranked. This is not one model's quirk: a fourth vendor fails the same threshold at 31.2% (Figure 7c). The taxonomy, and the finding that exemplars fix adherence but *not* invariance, are in Appendix L.

Inference-time repair does not survive the ceiling either. Self-consistency voting (Wang et al., 2023) is the obvious remedy, and ten replicates let us form $\tilde{I}$ under voting for the first time, since both sides of the ratio must themselves be voted. Voting raises raw cross-language agreement in both tested models but raises same-language agreement more, so on the ceiling-corrected estimand it costs 1.6 to 1.9 points with disjoint intervals: a variance reducer, not a retention improver. Most of our corrections make the effect larger (Table 1); this one and the scale extension of §4 instead remove results we could otherwise have claimed. We report both, because a correction pipeline that only ever helps is not one a reader should trust (Appendix O).

7 Conclusion

Answer agreement is not behavioural agreement. Given the same task in a different language, current tool-using models take a measurably different route, and under greedy decoding four frontier systems keep about the same share of their own action policy, whichever model it is. Below roughly 10B that regularity breaks, and the ordering among smaller models is largely an artifact of a chance floor we measured rather than assumed. The loss is structural, driven substantially by an English pivot the models will not abandon when instructed to and by a trace-length effect we establish causally; and much of what looks like multilingual failure is measurement failure, one regex having suppressed a competent model's measured accuracy twenty-sixfold. **An agent equally accurate in two languages is not thereby equally trustworthy in both**, because cost, failure modes and auditability live in the route rather than the answer, and making the action policy the measured object is what lets an evaluation see the difference.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267, 2023. doi: 10.18653/v1/2023.emnlp-main.258.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, 2020. doi: 10.18653/v1/2020.acl-main.421.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, 2024. doi: 10.18653/v1/2024.acl-long.44.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, 2018. doi: 10.18653/v1/D18-1269.

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

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York, 1993.

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

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, 2021. doi: 10.18653/v1/2021.findings-acl.413.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315–7330, 2020. doi: 10.18653/v1/2020.acl-main.653.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. arXiv:2211.09110.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. arXiv:2308.03688.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. arXiv:2311.12983.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, 2023. doi: 10.18653/v1/2023.acl-long.891.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Aaron Parisi, Yao Zhao, and Noah Fiedel. TALM: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2362–2376, 2020. doi: 10.18653/v1/2020.emnlp-main.185.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1994.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10650–10666, 2023. doi: 10.18653/v1/2023.emnlp-main.658.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024. arXiv:2307.16789.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 2022. doi: 10.1162/tacl_a_00452.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2302.04761.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design, or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. arXiv:2310.11324.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023. arXiv:2210.03057.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, 2024. doi: 10.18653/v1/2024.acl-long.820.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. arXiv:2210.03629.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020. doi: 10.18653/v1/2020.acl-main.148.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2402.18815.

Appendix

Appendix Guide

The appendices are grouped by what they are for. Method and estimator come first, then the data and the checks on it, then the full results behind each figure in the main text, then the boundary of the regularity, then housekeeping. Within each group the sections run in order.

	Appendix	What it contains
Reference	A Key terms	*rollout*, *cell*, *arm*, *confound* and the rest, defined once
Method	B Measurement protocol	the estimand, the five confounds, the controls
	C Estimator details	why the baseline is needed for identification, and the inference
	D Related work	where this sits against agentic and cross-lingual evaluation
	E The five confounds	each one priced on our own data
Data	F Data and coverage	sources, alignment keys, languages, the synthetic suite
	G Harness validity	routing, truncation and coverage, verified in the released data
	H Design checks	symmetry, seed, replicate agreement, subset representativeness
Results	I Per-cell results	all 48 matched cells with intervals
	J Trace length	what predicts irreproducibility, and the causal length manipulation
	K Uncorrected metrics	what a pipeline without the corrections would report
	L GPT-OSS failure taxonomy	the 26× accuracy artifact in full
	M Temperature ladder	all five rungs, with adherence
	N English-pivot ablation	four models, including the pre-registered head-room test
	O Self-consistency voting	the ten-replicate result
	P Correctness	accuracy by language, and why it is not a proxy
Boundary	Q The chance floor	the permutation null, and what correction changes
	R Scale and vendor	the three models outside the frontier four
Housekeeping	S Limitations	what remains open
	T Prompts	the scaffold and every intervention
	U Compute	hardware and cost
	V Provenance	the checks that fix which runs the numbers come from

A Key Terms

Several terms in Table 5 name units of the experimental design rather than standard usage, so they are fixed here once and used consistently throughout the paper.

Table 5: Key terms, fixed here and used consistently throughout the paper.

Term	Meaning in this paper
units of the experimental design	
rollout	One model run on one (task, language) pair. It yields one action trace and one final answer. All counts of the form "2.38M" are rollouts.
trace	The sequence of tool calls a rollout emits, for example `Translate`, `Search`, `Finish`. This, not the answer, is what we compare.
action policy	The mapping a model induces from a task to a trace. Because it is indexed by language, we can ask whether one model has one policy or several.
cell	One (model, benchmark, condition) unit, holding all the rollouts for that combination. "24 cells" means four models × six benchmarks.
arm	One experimental treatment applied to one model, for instance "`Translate` removed, Gemma-3-27B". An arm spans six cells.
replicate	An independent regeneration of a cell under a different serving seed, with everything else held fixed.
the estimator	
I_{within}, I_{cross}	Same-language and cross-language trace agreement, both measured between the two replicates.
gap Δ	$I_{\text{within}} - I_{\text{cross}}$: the raw shortfall, before any normalisation.
$\tilde{I}$	Normalised policy retention, $I_{\text{cross}} / I_{\text{within}}$: the share of a model's own reproducibility that survives a change of language.
ceiling-corrected	Expressed as $\tilde{I}$ rather than as Δ, so that a model's own irreproducibility is divided out instead of being read as a language effect.
frontier band	The $[0.708, 0.733]$ range of $\tilde{I}$ that the four frontier models occupy at $T{=}0$; "outside the band" is measured against it.

Term	Meaning in this paper
corrections and their objects	
confound	A property of the *measurement* rather than of the models that moves the reported number by as much as the effect being measured. We price five (Appendix E).
length matching	Reweighting the pairs being compared to a common trace-length distribution before averaging, done in both directions and accepted only when both agree in sign.
chance floor c	What the similarity score S returns for two traces that answer *different* tasks. Measured by permutation, not assumed (Appendix Q).
adherence	Whether a model emits a parseable trace at all. A model that does not is reported but excluded from retention comparisons.
parse failure	A rollout whose output contains no extractable `Action:` call. A subset are *empty traces*, where the output is blank; the rest name a tool in prose.
the mechanism	
English pivot	The habit of translating a non-English task into English before planning in it, visible as `Translate` opening a trace and as English-language reasoning text.
load-bearing	Said of the pivot: removing it degrades cross-lingual agreement, so it is doing causal work rather than being an incidental stylistic habit.
head-room	100 minus a model's baseline rate of opening a non-English rollout with `Translate`: how much room an instruction to pivot has to change behaviour.
pre-registered	Of a prediction: written into the analysis script, with the models chosen on a measured baseline property, before the compute was spent.

B Measurement Protocol

Let z be a *canonical task* and $x^{(\ell)}(z)$ its realisation in language ℓ, taken from a verified parallel alignment key rather than a row index (§2). Under a fixed scaffold with tool alphabet $\mathcal{T} = \{\texttt{Search}, \texttt{Calc}, \texttt{Translate}, \texttt{Summarize}, \texttt{Finish}\}$, a model emits an executed action trace $A^{(\ell)}(z) \in \mathcal{T}^*$ alongside its answer; the trace, not the answer, is what we compare. With $S(A, B) \in [0, 1]$ the normalised matching-block similarity, the naive corpus estimate is $I_{\text{cross}} = \mathbb{E}_{z,\ell_i \neq \ell_j} S(A^{(\ell_i)}, A^{(\ell_j)})$ (Appendix C).

That quantity has no meaningful zero, and its maximum is attainable by a degenerate policy: a model that always emits `Finish` scores 1.0. Worse, four confounds move it by as much as the effect being measured: a missing baseline, trace length, empty traces, and the reproducibility ceiling. Table 6 quantifies all four on this paper's own data, and three of them reverse a sign or a ranking there. A fifth, which we had to measure rather than correct, is the subject of the next paragraph.

The metric has no zero, so we measure it. S is a sequence-similarity score, and with a five-tool alphabet and short traces two *unrelated* policies already agree well above zero. Because $\tilde{I}$ is a ratio it is strictly more sensitive to an uncorrected floor than the difference is: under the standard model $S_{\text{obs}} = c + (1 - c)\, S_{\text{true}}$ the difference rescales cleanly and its ordering is invariant to c, whereas the ratio is biased toward 1 and differentially so by model. We therefore *measure* c rather than assume it, by permuting the task $\to$ trace assignment within a language arm and within length bin and recomputing S: the exact null "what does S score when two traces answer *different* tasks?" Across **66 cells** and 200 permutations, $\mathbf{c \approx 0.56}$, ranging from 0.35 on the longest-trace benchmark to **0.95** on the shortest. This is the difference between a hedge and a measurement, and §4 shows that it decides a headline ranking (Appendix Q).

We generate every cell **twice** under identical decoding configuration, task set and token budget, varying only the serving seed. Writing r_1, r_2 for the replicates,

$$I_{\text{within}} = \mathbb{E}_{z,\ell}\, S\big(A^{(\ell)}_{r_1}, A^{(\ell)}_{r_2}\big), \qquad I_{\text{cross}} = \mathbb{E}_{z,\ell_i \neq \ell_j}\, S\big(A^{(\ell_i)}_{r_1}, A^{(\ell_j)}_{r_2}\big). \tag{1}$$

Both sides are *cross-seed*, so decoding noise enters identically and language identity is the only difference. The **gap** is $\Delta = I_{\text{within}} - I_{\text{cross}}$ and **normalised policy retention** is $\tilde{I} = I_{\text{cross}} / I_{\text{within}}$: the share of a model's own reproducibility that survives a change of language.

Because S is length-sensitive (C2), every arm-to-arm comparison is reweighted to a common distribution over $\max(|A|, |B|)$ bins in both directions, and we accept a result only when both

agree in sign. Intervals are task-level bootstraps with 1,000 resamples (Efron & Tibshirani, 1993). Under C3 a pair is dropped whenever either trace is empty, which makes empty-trace and parse-failure rates first-class metrics and a finding in their own right (§6). Design checks are in Appendix H.

C Estimator Details

Appendix B states the estimator; this appendix supplies the properties it relies on, the identification argument, and the inference details.

Under a fixed scaffold with tool alphabet $\mathcal{T}$, a model θ induces $F_\theta(z, \ell) = (A^{(\ell)}(z),\, y^{(\ell)}(z))$, where $A^{(\ell)}(z) = \langle \tau_1, \ldots, \tau_k \rangle \in \mathcal{T}^*$ is the executed action trace and $y^{(\ell)}(z)$ the final answer. Outcome-scored evaluation observes only y; we observe A. We treat the induced behaviour as a family of language-indexed policies $\{\pi_\theta^{(\ell)}\}$ (Puterman, 1994) and ask how far apart its members are. Because $\mathcal{T}$ is fixed and language-independent, traces from two languages are sequences over the same alphabet and can be compared exactly, without a translation or entailment model in the measurement path.

The estimator has to be cross-seed on both sides. I_{within} and I_{cross} are both computed between replicate r_1 and replicate r_2, never within a single run. This symmetry is what a single-run estimate cannot offer: if I_{cross} were measured across languages within one run while I_{within} compared two runs, the two quantities would differ in *two* respects (language and seed) rather than one, and the difference could not be attributed to language. Under the cross-seed construction decoding noise enters both sides identically and language identity is the only remaining difference.

The similarity $S(A, B)$ is the normalised matching-block similarity (the ratio $2M/(|A| + |B|)$, where M is the total size of matched blocks under a longest- contiguous-match decomposition). It is symmetric, equals 1 for identical sequences and 0 for sequences sharing no tool. Two properties drive the confounds. It has **no meaningful zero**: with $|\mathcal{T}| = 5$ and short traces, two unrelated policies already score well above zero. And it is **length-sensitive**: for a fixed edit rate, shorter sequences score higher, which is why every arm-to-arm comparison in this paper is length-matched (C2).

I_{within} is not ≈ 1, and without it the quantity is not identifiable. A model re-run on the *same* prompt in the *same* language under a different serving seed does not reproduce its own trace: we measure $I_{\text{within}} = 0.63$–0.80 depending on model and temperature. This is the identification problem in one line. A cross-language observation $S(A^{(\ell_i)}, A^{(\ell_j)})$ is depressed by two sources at once, the model's own stochasticity and the change of language, and a single run supplies one equation in two unknowns. Any number computed from it is a language effect only under the assumption $I_{\text{within}} = 1$, which is false by 20–37 points here and false by a *model-dependent* amount, so the assumption does not even preserve orderings (§4). The matched replicate supplies the second equation, and it must be measured rather than borrowed: reusing an existing run introduces a token-budget mismatch that alone moves the headline from $+0.0861$ to $+0.0625$, because a smaller budget truncates traces, which shortens them, which inflates S on both sides but not equally.

Intervals are task-level bootstraps with 1,000 resamples (Efron & Tibshirani, 1993). Resampling is over *tasks* rather than over pairs, because the $\binom{L}{2}$ language pairs generated by one task are not independent; resampling pairs would understate the intervals substantially. The same resample indices are used for I_{within} and I_{cross} within a draw, so $\tilde{I}$ and Δ inherit the correlation between numerator and denominator rather than treating them as independent.

Under C3 a pair is dropped whenever *either* trace is empty. The strict rule matters: retaining empty-vs-empty pairs is what lifts GPT-OSS-120B to the top of the raw invariance ranking (Appendix L), and the correction is worth up to $-\mathbf{0.119}$ on a single cell. Because exclusion discards information non-uniformly across languages, empty-trace and parse-failure rates are reported as first-class metrics everywhere rather than being silently absorbed into the invariance number.

The symmetry correction, a same-seed control and subset representativeness are verified in Appendix H.

D Extended Related Work

ReAct-style scaffolds (Yao et al., 2023) and learned tool use (Schick et al., 2023; Parisi et al., 2022; Qin et al., 2024) established the interleaved reason–act format we adopt: the model alternates free-form deliberation with a symbolic call drawn from a declared tool inventory. Agentic benchmarks evaluate task completion (Liu et al., 2024; Mialon et al., 2024) but are overwhelmingly English and score *outcomes*, using the action sequence only as a means to that score. Two consequences follow for our purposes. First, a model that reaches the right answer by a different route is indistinguishable from one that reaches it by the intended route, so behavioural variation is invisible by construction. Second, because scoring depends on extracting a final answer, these harnesses inherit a parsing step whose failure modes are not audited; §6 and Appendix L show that this step alone can suppress a model's measured accuracy by more than an order of magnitude. We instead make the trace the measured object and ask a comparative question about it across languages, which requires a same-language baseline that outcome-scored suites do not need and do not collect.

Cross-lingual suites (Hu et al., 2020; Ahuja et al., 2023; Liang et al., 2023) compare final predictions, and the underlying datasets we build on (Conneau et al., 2018; Ponti et al., 2020; Artetxe et al., 2020; Bandarkar et al., 2024) are translations or careful adaptations of a common item pool. Their alignment guarantees are exactly what makes a matched cross-lingual comparison possible, but they are guarantees about *items*, not about pipelines: Appendix F documents five multilingual corpora whose per-language splits do not correspond item-for-item despite being distributed together, and one (Belebele) that requires re-keying before an index join is valid.

Chain-of-thought work and its multilingual extension (Wei et al., 2022; Shi et al., 2023) and multitask finetuning (Muennighoff et al., 2023) examine intermediate *text*. Text is the natural object when the question is whether a model reasons at all, but it is a poor object for a cross-lingual comparison: two rationales in different languages cannot be scored for agreement without a translation or entailment model, which introduces its own multilingual error profile into the measurement. Our shared symbolic action alphabet avoids this entirely. The comparison is between sequences over a fixed, language-independent vocabulary of tool names, so agreement is computed exactly rather than estimated.

Closest in spirit, Qi et al. (2023) measure whether models retrieve the same *facts* across languages and report substantial inconsistency. We ask the analogous question about *procedures*. The distinction matters because a procedure is chosen rather than recalled: the model can reach the same answer through different tools, so procedural divergence is not simply knowledge divergence observed at another layer. It also changes what must be controlled. A fact is either retrieved or not, whereas a procedure is sampled, so any cross-lingual procedural comparison confounds language with decoding noise unless a matched same-language baseline is measured under identical conditions. Appendix C shows that without this baseline the quantity is not identifiable, and §4 shows that the uncorrected version ranks models by their determinism rather than by their cross-lingual behaviour.

Wendler et al. (2024) and Zhao et al. (2024) give representational evidence that multilingual transformers route non-English inputs through English-aligned internal states. That evidence is correlational with respect to behaviour: it establishes that an English-like representation is present, not that downstream behaviour depends on it. Our ablation is the behavioural, causal counterpart, removing the pivot from the action space and observing the effect on cross-lingual agreement (§5, Appendix N). The two directions of evidence agree, and the behavioural version adds a result the representational version cannot supply: the pivot survives a direct instruction not to use it, at a $>$99% refusal rate in both models tested.

Sclar et al. (2024) show that input-formatting choices move benchmark scores substantially, and that this sensitivity is large enough to reorder models. §6 documents the output-side

analogue: a single-pattern trace extractor, applied uniformly to all models and languages, suppressed one model's measured accuracy by 26× and simultaneously promoted it to the top of the raw invariance ranking. The output-side case is arguably the more dangerous of the two, because a formatting choice is visible in the prompt whereas an extractor lives in the harness, and because parse failure need not be language-uniform, so the artifact can present as a multilingual finding.

E The Five Confounds

Every result in the paper corrects for all five confounds below, and each is large enough on its own to change a published conclusion. Four do more than that on this paper's own data: C2 *reverses the sign* of two interventions, C3 and C4 *reverse a model ranking* (C3 by demoting the model whose empty pairs score 1.0, C4 by ranking determinism rather than language, $\rho = -0.80$ between two temperatures), and C5 *reverses* which of two smaller models retains more. C1's cost is a 38% underestimate of the headline rather than a reversal, which is precisely why it went unnoticed. Impacts are measured on this paper's own data. Two of the five are established *causally* rather than by correction alone: C2 by a three-level manipulation (Appendix J.1) and C5 by a permutation null over 66 cells (Appendix Q).
C5, the chance floor: S has no meaningful zero, because unrelated policies over a five-tool alphabet already agree over half the time; measured $c \approx 0.56$ and up to 0.95 on the shortest traces, and correcting for it *reverses* which of two smaller models shows higher retention.

Confound	Symptom if ignored	Correction	Measured impact
C1 no baseline	Decoding noise is scored as a language effect	Generate every cell twice; identical budget, seed is the only difference	$I_{\text{within}} = 0.63$–0.80, not ≈ 1; matching the budget makes the estimate **38% larger**
C2 trace length	Short traces score higher, so "hard benchmark" and "long trace" look alike	Reweight to a common max-length bin distribution, in *both* directions	$R^2 = 0.67$ across cells (0.92 within one model); flips the *sign* of two interventions
C3 empty traces	Two empty traces are maximally similar, so non-adherence is rewarded	Drop a pair when *either* side is empty; report the empty rate separately	Up to $-\mathbf{0.119}$ on one cell; collapses one model's headline invariance $0.79 \to 0.02$
C4 ceiling	The gap is bounded by a model's own reproducibility, so rankings rank determinism	Normalise: $\tilde{I} = I_{\text{cross}} / I_{\text{within}}$	$r(I_{\text{within}}, \text{gap}) = +\mathbf{0.97}$ at $T = 0$; cross-model spread $81\% \to 34\%$
C5 chance floor	Unrelated traces already agree by construction, so S has no meaningful zero	Measure the floor by permuting task→trace within a language arm, within length bin	$c \approx \mathbf{0.56}$, up to 0.95 on short traces; *reverses* which of two smaller models retains more

Table 6: The five confounds in cross-lingual policy measurement, each quantified on this paper's own data. C2 and C5 are additionally established causally, by manipulation and by permutation null respectively.

C1 dominates the others, and the reason is structural: they are metric artifacts, whereas C1 is an identification failure. Without a same-language baseline, I_{cross} has no reference point at all, and the natural fallback of reusing an existing single run silently introduces a token-budget mismatch. On our data that mismatch alone moves the headline from $+0.0861$ to $+0.0625$, and it moves it in the *conservative* direction, which is why the confound went unnoticed until a matched baseline was generated.

The corrections are not commutative in magnitude, though they are in sign. We apply C1 (matched replicates) at generation time, then C3 (strict empty exclusion) at pair construction, then C2 (length matching) at aggregation, then C4 (normalisation) at reporting. Applying C2 before C3 changes the pooled estimate by < 0.004 because empty pairs sit in the shortest length bin and are dropped either way.

F Data and Coverage

This appendix documents where the tasks come from, what guarantees their cross-lingual correspondence, and what had to be rebuilt before they could be used. The organising constraint is stated in Appendix C: a cross-lingual comparison of *procedures* is only well-posed if $x^{(\ell_i)}(z)$ and $x^{(\ell_j)}(z)$ are the same underlying task z. That is a property of the

alignment key, not of the corpus name, so we record the key used for every benchmark below and treat any dataset without one as unusable rather than as noisy.

Benchmark	Source	Tasks	Langs	Rollouts	Pairs	Alignment key
FLORES-200	NLLB Team et al. (2022)	997	21	20,937	209,370	sentence `id`
XQuAD	Artetxe et al. (2020)	1,189	12	14,268	78,474	question `id`
XNLI	Conneau et al. (2018)	2,490	15	37,350	261,450	`all_languages` config
Belebele	Bandarkar et al. (2024)	900	16	14,400	108,000	`link+question_no.`
XCOPA	Ponti et al. (2020)	100	11	1,100	5,500	`idx`
Synthetic	this work	100	23	2,300	25,300	construction
Total		**5,776**	**41**†	**90,355**	**688,094**	

Table 7: Evaluation suite, *per model*. †41 distinct languages across the suite, not a sum. Gold answers exist for XQuAD, XNLI, Belebele and XCOPA; FLORES-200 has none by construction. XCOPA is the only benchmark with no English arm. The orange cell is the one key we had to construct ourselves: Belebele ships no usable index, and joining on row order matches only 51% of items.

F.1 The experimental campaign

Table 8 lists every condition generated for this paper in one place, so that the scale claim in §2 can be audited row by row rather than taken on trust. Rows are ordered as the paper uses them: the main sweep and the matched replicates that make the estimand identifiable, then the conditions that remove sampling, then the six interventions, then the CPU-only permutation null. Every row is generated under the same harness and the same decoding configuration except where the row is itself the manipulation.

F.2 Item-level parallelism

Item-level parallelism is a property of the alignment key, not of the corpus name, and it is scarcer than it looks. Screening candidate corpora against it rules out Samanantar (Ramesh et al., 2022) and OPUS-100 (Zhang et al., 2020), which are En–X bitext with no X–Y correspondence; XL-Sum (Hasan et al., 2021), whose articles are collected independently per language; MLQA (Lewis et al., 2020), which is parallel only through a shared question id and whose 7 languages are a strict subset of XQuAD's 12; and TyDi QA (Clark et al., 2020), which is not parallel by design. A pipeline that indexes any of them by row runs without error and returns a cross-lingual invariance number that means nothing, which is the failure mode this appendix exists to rule out for the six benchmarks we do use.

F.3 Building the suite

The six benchmarks were not usable as distributed either. Each defect below is an ordinary data-engineering fault rather than anything exotic, which is precisely why it is easy to inherit silently: every one yields a corpus that loads, iterates and scores without raising an error.

F.4 Languages

Forty-one languages, 17 families, 16 writing systems, 809 distinct language pairs. Resource tier follows the standard high/medium/low split (23/11/7). The suite is deliberately unbalanced toward Indic languages, which is what makes the adherence-collapse result of §6 visible: a suite of only high-resource European languages would not have exposed it.

Bodo, Dogri and Konkani appear only in the synthetic benchmark; no parallel benchmark reaches them. Script compliance of the released files is 97.1–100% per language, measured by Unicode block membership.

Condition	Purpose	Models	Cells	Rollouts
Main sweep (T=0.5)	Raw metrics, adherence	5	30	451,775
Matched replicates (T=0.5)	Baseline I_{within} (C1)	5	30	226,000
Greedy (T=0)	Remove sampling noise	4	24	180,800
Temperature ladder	$T \in \{0, 0.3, 0.5, 0.7, 1.0\}$	2	30	113,000
Few-shot exemplars	Adherence repair	2	24	90,400
Tool-alphabet ablation	Pivot mechanism (causal)	2	24	90,400
Reasoning-language control	Pivot mechanism (causal)	2	19	74,100
Pivot head-room *(pre-registered)*	Mechanism, 4 models	2	36	135,600
Voting, 5 replicates	Inference-time repair	2	60	162,000
Voting, 10 replicates ($k \leq 5$)	Ceiling-corrected repair	2	120	452,000
Scale & vendor generalisation	Boundary of the regime	3	36	135,600
Trace-length dose–response	C2 causality	2	72	271,200
Chance floor (permutation)	C5, CPU-only	8	*(66)*[†]	—
Total		**8**	**505**	**2,382,875**

Table 8: The experimental campaign. A *cell* is one (model, benchmark, condition) unit. [†]The chance-floor analysis is a permutation null recomputed over cells that the rows above already generated, so it contributes neither rollouts nor cells of its own; its 66 are parenthesised and excluded from both totals, and the twelve generated rows sum exactly to 505 and 2,382,875. Ablation arms use a 300-task subset per benchmark, which reproduces full-sweep invariance to within ± 0.02 (Appendix H).

Issue in the distributed data	What we did	Effect
Belebele carries no usable cross-language index	Re-keyed on `link` + `question_no.`; a row-order join matches 51% of items, with gold agreeing on 64%	900/900 items aligned, gold 100%
Only a subset of each corpus's languages is exposed by default	Rebuilt every benchmark at full language coverage	XNLI 3→15, XQuAD 3→12, FLORES 12→21, Belebele 11→16
Gold answers are not carried through to the task files	Gold emitted at build time	4 benchmarks, 4,679 tasks
Some synthetic instances came back untranslated	Regenerated	515 of 2,200 (23.4%) → 0
Some synthetic instances came back romanised	Regenerated in native script	678 instances, 97.5% compliant

Table 9: What the suite required before any measurement. The two blue cells matter most: a 51% item-match rate silently compares different questions across languages, and untranslated instances make a "non-English" arm partly English.

Code	Language	Family	Script	Code	Language	Family	Script
ar	Arabic	Semitic	Arabic	mni	Manipuri	Sino-Tibetan	Bengali
as	Assamese	Indo-Aryan	Bengali	mr	Marathi	Indo-Aryan	Devanagari
bg	Bulgarian	Slavic	Cyrillic	ne	Nepali	Indo-Aryan	Devanagari
bn	Bengali	Indo-Aryan	Bengali	or	Odia	Indo-Aryan	Oriya
brx	Bodo	Sino-Tibetan	Devanagari	pa	Punjabi	Indo-Aryan	Gurmukhi
de	German	Germanic	Latin	qu	Quechua	Quechuan	Latin
doi	Dogri	Indo-Aryan	Devanagari	ro	Romanian	Romance	Latin
el	Greek	Hellenic	Greek	ru	Russian	Slavic	Cyrillic
en	English	Germanic	Latin	sa	Sanskrit	Indo-Aryan	Devanagari
es	Spanish	Romance	Latin	sat	Santali	Austroasiatic	Ol Chiki
et	Estonian	Uralic	Latin	sd	Sindhi	Indo-Aryan	Arabic
fr	French	Romance	Latin	sw	Swahili	Bantu	Latin
gom	Konkani	Indo-Aryan	Devanagari	ta	Tamil	Dravidian	Tamil
gu	Gujarati	Indo-Aryan	Gujarati	te	Telugu	Dravidian	Telugu
hi	Hindi	Indo-Aryan	Devanagari	th	Thai	Kra-Dai	Thai
ht	Haitian	French Creole	Latin	tr	Turkish	Turkic	Latin
id	Indonesian	Austronesian	Latin	ur	Urdu	Indo-Aryan	Arabic
it	Italian	Romance	Latin	vi	Vietnamese	Austroasiatic	Latin
kn	Kannada	Dravidian	Kannada	zh	Chinese	Sinitic	Han
ks	Kashmiri	Indo-Aryan	Arabic	mai	Maithili	Indo-Aryan	Devanagari
ml	Malayalam	Dravidian	Malayalam				

F.5 Synthetic benchmark composition

The suite was generated and translated with `gpt-5-mini` (deployment `2025-08-07`) at its default decoding settings; the generation prompt is in Appendix T. No model evaluated in this paper was used to author it. 100 tasks × 23 languages. Task families: constraint planning

(18), retrieval–calculation pipeline (17), evidence reconciliation (17), policy-compliance exception (16), schedule/resource allocation (16), error-audit correction (16). Difficulty: medium (51), hard (30), easy (19). Each task carries `required_tools`, `expected_min_steps` and `complexity_tags`. Median prompt length is 3,233 characters, against 139–800 for the adapted benchmarks, which is why its traces are 1.7–4.6× longer and why it must never be compared to the adapted benchmarks without length matching.

G Harness Validity

Three properties are verified in the delivered data rather than assumed, because each can fail silently: a harness that routes every language arm to the same prompt, truncates long traces, or drops rollouts from a subset of arms will still produce a complete-looking results table. The checks below are run on the released data, not on a separate test fixture.

Language routing is measured from each run's own prompt log on Belebele, over 16 arms and 900 tasks. Every arm holds 900 distinct prompts, **prompt sharing between arms is 0.0%** for every model, and mean ASCII is 0.348 with exactly one Latin-script arm, English. A run that routes every arm to the same text shows 100% sharing at 0.999 ASCII on every arm; ours shows none.

Truncation, meaning `finish_reason == length` at `max_tokens`=4096, is 0.003% for Gemma, 0.005% for Sarvam, 0.004% for Llama-4, 0.61% for Qwen3 and 0.63% for GPT-OSS, so no result in this paper is budget-limited. The field `used_reasoning_channel` is 0 on all 451,775 main-sweep rollouts for every model, which excludes by direct measurement rather than by argument the hypothesis that GPT-OSS's short outputs reflect a separate reasoning channel the harness discarded (§6). And coverage is exact: every cell matches tasks × languages, for 451,775 of 451,775 rollouts, with no ragged language arm and no missing task.

H Design Checks

The symmetry correction is not an artifact. Our estimator takes I_{cross} across replicates. Compared against the same-language-run convention, the two agree to within 0.002 on every model (Gemma 0.6728/0.6734, Sarvam 0.5855/0.5878, Qwen3 0.4991/0.4977), so the choice does not drive any result.

Serve seed is irrelevant at $T = 0$, as it must be. Two complete, independent greedy runs exist at the *same* seed. Comparing them isolates batching non-determinism with everything else fixed: Gemma I_{within} 0.9047 vs. 0.9060, length-matched gap +0.2110 vs. +0.2131; Sarvam 0.8604 vs. 0.8632 and +0.2200 vs. +0.2224. Two jobs launched hours apart on different hardware agree to three decimal places.

Replicate agreement under greedy decoding is exact where it should be. Llama-4's two greedy replicates produce *identical* empty-trace counts (173/173 on the synthetic benchmark) and mean trace lengths matching to two decimals on every benchmark.

The 300-task subset is representative. Ablation arms use 300 tasks per benchmark rather than the full sweep. Invariance computed on the subset reproduces full-sweep invariance to within ±0.02.

Benchmarks are resolved by content, not by name. Run directory names are length-capped and can silently lose the benchmark suffix for long model identifiers, so all analyses identify a benchmark by its *language-count fingerprint* (21/12/15/16/11/23), which is unique across the suite and survives any renaming. Main-sweep invariance recomputes to the reported values exactly ($\Delta = 0.0000$) under this resolver.

I Per-Cell Results

The pooled numbers in the main text average over cells of very different sizes, from XCOPA's 1,100 same-language pairs to FLORES-200's 63,000 cross-language ones. This appendix

therefore reports every cell separately, so that a reader can check whether any single benchmark or model carries the result. It does not: all 48 length-matched gaps are positive, and at $T{=}0$ every bootstrap interval on the raw gap excludes zero.

Figure 4 in the main text plots these cells; the numeric values, bootstrap intervals and length-matched counterparts are tabulated below, under strict empty exclusion, a symmetric cross-seed contrast and a 1,000-sample task-level bootstrap.

Every headline quantity in this paper is computed on *pooled pairs*, which weights a cell by how many language pairs it contributes. Table 2 instead prints I_{within} and I_{cross} macro-averaged over the six benchmarks, so a reader who divides those two columns will not exactly recover the printed $\tilde{I}$. We state both here so the difference is visible rather than discovered. Recomputed from the printed model means, $\tilde{I}$ at $T{=}0$ is 0.746 (Gemma), 0.715 (Sarvam), 0.722 (Qwen3) and 0.713 (Llama-4), a span of 3.2 points against the pooled 2.6; the ceiling correlation is $+0.51$ at $T{=}0.5$ and $+0.94$ at $T{=}0$ against the pooled $+0.43$ and $+0.97$; and the absolute-gap relative spread falls 87% $\rightarrow$ 44% against the pooled 81% $\rightarrow$ 34%. No conclusion in the paper depends on which aggregation is used: the ordering, the direction and the collapse at $T{=}0$ are identical under both.

Dividing the two invariance columns of the $T{=}0$ block below gives $\tilde{I}$ for each of the 24 cells, which is how §4 answers the $n = 4$ objection (Table 3). Because the division happens *within* a cell, trace length enters numerator and denominator identically and largely cancels, making this the only benchmark-level comparison in the paper that length matching is not required to interpret.

J Trace Length and Irreproducibility

Qwen3-235B is a genuine outlier, retaining 32% within-language dissimilarity even at $T{=}0$ against 9–14% for the other three. On three models an attractive explanation is batch-dependent expert routing in mixture-of-experts serving. **Llama-4 refutes it**: it is a 128-expert MoE and the *second most* reproducible model in the study ($1 - I_{\text{within}} = 0.0995$), statistically indistinguishable from dense Gemma (0.0940) and better than dense Sarvam (0.1368). We report the hypothesis alongside its refutation because on three models it looks publishable and is wrong.

Testing candidate correlates *per cell* across all 24 greedy cells, rather than across four model means:

J.1 The three-level length manipulation

Everything above is correlational, and the correlation is strong enough ($R^2 = 0.67$ across the 24 adherent cells, 0.92 within Gemma alone; Figure 8a) that a difficulty account and a length account are observationally equivalent. We therefore manipulated length directly. Three prompt-level arms (*at most 2* actions, *at least 5*, and *at least 8*) were run at $T{=}0$ on two models with task, language, seed, decoding and token budget held fixed, for 72 cells and 271,200 rollouts. Three levels rather than two matters: a monotone move across three points is hard to attribute to an uncontrolled covariate, and it also lets a non-monotone response be *seen* rather than averaged away.

The manipulation took. Mean actions per rollout moves roughly threefold in both models, Gemma 2.86 $\rightarrow$ 6.32 $\rightarrow$ 9.09 and Qwen3 3.45 $\rightarrow$ 5.67 $\rightarrow$ 8.68, monotonically and in the instructed direction. Empty traces are 0% in every cell except Qwen3's synthetic (21%, the pre-existing low-resource adherence collapse of §6), so the arms are not confounded with adherence.

Two things follow, and the second overturns a natural assumption. First, in Qwen3 the response is textbook, with longer traces, a larger raw gap and lower retention, monotone on all three points, so the $R^2 = 0.67$ correlation is at least partly causal rather than a difficulty proxy. Second, and against a natural intuition, **the normalisation does not absorb length**. Because $\tilde{I}$ divides I_{cross} by I_{within} *within* a cell, it is natural to argue that length enters

Model	Benchmark	n_w	n_c	I_{within}	I_{cross}	Raw gap	Len-match	95% CI (raw gap)
$T = 0.5$, *matched replicates*								
Gemma-3-27B	FLORES-200	6,300	63,000	0.7327	0.6434	0.0894	0.0669	[+0.0836, +0.0946]
	XQuAD	3,600	19,800	0.8626	0.7379	0.1248	0.0994	[+0.1143, +0.1362]
	XNLI	4,500	31,500	0.8980	0.7439	0.1542	0.1450	[+0.1476, +0.1600]
	Belebele	4,800	36,000	0.8141	0.6936	0.1205	0.1078	[+0.1126, +0.1287]
	XCOPA	1,100	5,500	0.9018	0.7433	0.1585	0.1422	[+0.1463, +0.1703]
	Synthetic	2,291	25,121	0.5731	0.4783	0.0948	0.0822	[+0.0847, +0.1050]
Sarvam-M	FLORES-200	6,237	62,261	0.5985	0.5669	0.0316	0.0266	[+0.0271, +0.0360]
	XQuAD	3,513	19,268	0.7364	0.6935	0.0428	0.0353	[+0.0349, +0.0513]
	XNLI	4,396	30,684	0.6079	0.5543	0.0535	0.0421	[+0.0458, +0.0610]
	Belebele	4,665	35,000	0.7727	0.7397	0.0330	0.0220	[+0.0259, +0.0405]
	XCOPA	1,082	5,393	0.6305	0.5604	0.0702	0.0617	[+0.0585, +0.0833]
	Synthetic	1,859	18,660	0.4362	0.4122	0.0239	0.0220	[+0.0154, +0.0328]
Qwen3-235B	FLORES-200	6,298	62,971	0.7028	0.5637	0.1390	0.1050	[+0.1303, +0.1480]
	XQuAD	3,600	19,800	0.6051	0.4215	0.1836	0.1499	[+0.1719, +0.1947]
	XNLI	4,500	31,500	0.6577	0.4885	0.1691	0.1354	[+0.1589, +0.1797]
	Belebele	4,799	35,987	0.6372	0.5259	0.1112	0.0907	[+0.1027, +0.1193]
	XCOPA	1,100	5,500	0.7131	0.6079	0.1052	0.0865	[+0.0903, +0.1210]
	Synthetic	1,822	16,978	0.5091	0.3784	0.1307	0.0933	[+0.1126, +0.1493]
Llama-4-Mav.	FLORES-200	6,300	63,000	0.7474	0.6458	0.1016	0.0879	[+0.0967, +0.1066]
	XQuAD	3,600	19,800	0.7978	0.6612	0.1366	0.1210	[+0.1280, +0.1452]
	XNLI	4,500	31,500	0.7835	0.6896	0.0939	0.0840	[+0.0873, +0.1004]
	Belebele	4,798	35,986	0.8070	0.7066	0.1003	0.0892	[+0.0948, +0.1066]
	XCOPA	1,100	5,500	0.7638	0.6580	0.1058	0.0964	[+0.0912, +0.1203]
	Synthetic	2,086	21,695	0.5910	0.5113	0.0797	0.0746	[+0.0709, +0.0885]
$T = 0$, *greedy*								
Gemma-3-27B	FLORES-200	6,300	63,000	0.8749	0.6452	0.2297	0.1924	[+0.2209, +0.2383]
	XQuAD	3,600	19,800	0.9632	0.7432	0.2200	0.2049	[+0.2061, +0.2339]
	XNLI	4,500	31,500	0.9651	0.7440	0.2211	0.2150	[+0.2138, +0.2284]
	Belebele	4,800	36,000	0.9520	0.6948	0.2572	0.2419	[+0.2471, +0.2678]
	XCOPA	1,100	5,500	0.9745	0.7479	0.2266	0.2183	[+0.2138, +0.2385]
	Synthetic	2,292	25,126	0.7062	0.4776	0.2287	0.2057	[+0.2139, +0.2436]
Sarvam-M	FLORES-200	6,267	62,436	0.8505	0.5741	0.2764	0.2355	[+0.2705, +0.2821]
	XQuAD	3,561	19,465	0.9399	0.7292	0.2107	0.1932	[+0.1972, +0.2229]
	XNLI	4,459	31,059	0.8883	0.5850	0.3033	0.2809	[+0.2946, +0.3119]
	Belebele	4,729	35,254	0.9625	0.7846	0.1779	0.1622	[+0.1663, +0.1900]
	XCOPA	1,097	5,471	0.8932	0.5779	0.3153	0.2994	[+0.3033, +0.3280]
	Synthetic	2,027	19,997	0.6446	0.4543	0.1903	0.1632	[+0.1751, +0.2061]
Qwen3-235B	FLORES-200	6,300	63,000	0.7400	0.5591	0.1809	0.1392	[+0.1716, +0.1903]
	XQuAD	3,600	19,800	0.6661	0.4093	0.2568	0.2148	[+0.2441, +0.2699]
	XNLI	4,499	31,490	0.6979	0.4815	0.2163	0.1750	[+0.2063, +0.2267]
	Belebele	4,800	36,000	0.6683	0.5307	0.1376	0.1149	[+0.1291, +0.1467]
	XCOPA	1,100	5,500	0.7488	0.5814	0.1674	0.1385	[+0.1522, +0.1837]
	Synthetic	1,820	16,921	0.5510	0.3789	0.1721	0.1280	[+0.1514, +0.1923]
Llama-4-Mav.	FLORES-200	6,300	63,000	0.9369	0.6394	0.2975	0.2800	[+0.2910, +0.3036]
	XQuAD	3,600	19,800	0.9204	0.6557	0.2647	0.2459	[+0.2553, +0.2751]
	XNLI	4,500	31,500	0.9417	0.6877	0.2540	0.2410	[+0.2467, +0.2609]
	Belebele	4,799	35,985	0.9273	0.7108	0.2165	0.1999	[+0.2092, +0.2233]
	XCOPA	1,100	5,500	0.9490	0.6557	0.2933	0.2826	[+0.2777, +0.3086]
	Synthetic	2,106	21,750	0.7277	0.5052	0.2226	0.2059	[+0.2072, +0.2376]

Table 10: All 48 matched cells. Every gap is positive, and at T=0 every interval excludes zero. **The intervals are task-level bootstraps on the raw gap, not on the length-matched value**, so the length-matched column is not bracketed by them; since the raw gap exceeds the length-matched gap in every cell and both are positive, the exclusion of zero carries to both.

Predictor of $1 - I_{within}$	r	R^2
Mean trace length	+0.799	**0.639**
Mean agent iterations	+0.665	0.442
Collinearity (iterations ↔ length)	+0.425	—
Mixture-of-experts indicator	—	no signal

numerator and denominator together and cancels. It does not. The ratio moves 6.0 and 6.9 points with disjoint intervals, which is *larger* than the entire across-model band the paper's central result rests on (Figure 5c). Every benchmark-level statement that assumed $\tilde{I}$ was length-neutral is withdrawn, and length is treated throughout as a first-order driver rather than a nuisance parameter.

The sign is model-dependent, so no universal claim is made. Gemma is non-monotone in both $\tilde{I}$ and the raw gap, and the anomaly sits at the *short* end: constraining it to at most two actions gives it the lowest retention of the three arms (0.7689 against 0.8292 at ≥ 5).

Model	Arch.	Mean iter. (range)	Trace len. (range)	$1 - I_{\text{within}}$ (range)
Gemma-3-27B	dense	1.11–3.40	1.65–11.67	0.026–0.294
Llama-4-Maverick	MoE	1.00–2.19	2.37–7.64	0.051–0.272
Sarvam-M	dense	1.03–2.99	1.41–10.92	0.038–0.355
Qwen3-235B	MoE	2.42–4.72	4.24–9.53	0.251–0.449

Table 11: Per-cell predictors of greedy irreproducibility. Trace length is the strongest single correlate and architecture carries no signal. The table also records a hypothesis the fourth model rules out: on three models, Qwen3's outlier status is naturally attributed to batch-dependent expert routing in MoE serving. Llama-4 refutes it: a 128-expert MoE that is the *second most* reproducible model in the study. At $n = 4$ model means the ordering of length and iterations reverses, which is a small-sample artifact and a reminder that model-level correlations on a handful of systems are not evidence.

Model	Arm	Mean actions	I_{within}	I_{cross}	$\tilde{I}$	95% CI
Gemma-3-27B	≤ 2 actions	2.86	0.9380	0.7212	0.7689	[0.7632, 0.7751]
	≥ 5 actions	6.32	0.8817	0.7311	0.8292	[0.8233, 0.8348]
	≥ 8 actions	9.09	0.8240	0.6696	0.8126	[0.8070, 0.8179]
Qwen3-235B	≤ 2 actions	3.45	0.7940	0.6414	0.8078	[0.8032, 0.8125]
	≥ 5 actions	5.67	0.9813	0.7578	0.7722	[0.7644, 0.7799]
	≥ 8 actions	8.68	0.9653	0.7136	0.7392	[0.7325, 0.7454]

Table 12: The three-level trace-length dose–response, at T=0 on 300 tasks with two replicates per arm. The blue cells are the extreme arms, whose intervals are disjoint in both models: $\tilde{I}$ moves 6.0 points in Gemma and 6.9 in Qwen3 under a 3× change in length, against a 2.6-point band across the four frontier models. Qwen3 is monotone decreasing on all three points; Gemma is not, and its anomaly is the *shortest* arm.

One reading, offered as a hypothesis rather than a result, is that at two actions a model has no room to recover from a divergent opening move, whereas by five it has slack to reconverge; that predicts a minimum at the short end, which is what Gemma shows and which Qwen3, whose shortest arm already averages 3.45 actions, may simply never reach. Two models and three points cannot establish a U-shape. What is established is that the effect is **model-dependent in sign**, which by itself defeats any universal statement about length. Per benchmark (Figure 8c), Belebele is the one cell monotone decreasing in *both* models (0.818 → 0.765 and 0.868 → 0.710) and is the cleanest single instance of the causal effect.

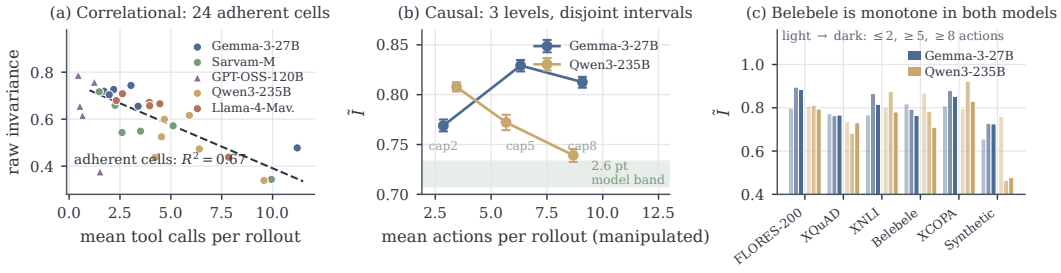

Figure 8: Trace length, from correlation to cause. (a) The correlational picture across all 30 main-sweep cells, with the fit taken on the 24 protocol-adherent ones; triangles are GPT-OSS, excluded. (b) The three-level manipulation, pooled, with the 2.6-point across-model band shaded for scale: $\tilde{I}$ moves further under length than across four frontier models. (c) Per benchmark, the sign disagrees between models everywhere except Belebele.

Four caveats travel with this. The instruction is a *soft* constraint, since "at most 2" and "at least 5" are prompt text rather than a decoding limit, so compliance is partial and the arms overlap in the tails; the realised means in Table 12 are what the manipulation actually achieved. Three levels is the minimum at which monotonicity is testable, and Gemma fails it, so a four- or five-level ladder is needed to separate a U-shape from noise. Two models cannot settle the sign. And every capped arm sits above its own uncapped published baseline, which must *not* be read as "capping improves retention": these arms use the

300-task subset while the published baselines use the full sweep, and the across-benchmark spread within a single model reaches 14 points. The arms are comparable *to each other*, which is what the design requires.

K Uncorrected Metrics

These are the numbers a pipeline without the Appendix E corrections would report. They are included so that the size of each correction is auditable, **not** as results.

Model	FLORES	XQuAD	XNLI	Belebele	XCOPA	Synthetic
Raw invariance (no baseline, no length match, no empty exclusion)						
Gemma-3-27B	0.6550	0.7189	0.7265	0.7040	0.7435	0.4777
Sarvam-M	0.5717	0.6590	0.5433	0.7168	0.5490	0.3437
GPT-OSS-120B	0.3757	0.6152	**0.7920**	0.6549	**0.7866**	0.7572
Qwen3-235B	0.5997	0.4368	0.4732	0.5248	0.6162	0.3391
Llama-4-Maverick	0.6654	0.6791	0.6717	0.7083	0.6571	0.4383
Parse-failure rate						
Gemma-3-27B	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%
Sarvam-M	0.7%	1.4%	1.5%	1.5%	0.6%	14.7%
GPT-OSS-120B	51.1%	75.5%	**88.6%**	79.9%	**88.5%**	85.7%
Qwen3-235B	0.0%	0.0%	0.0%	0.0%	0.0%	20.7%
Llama-4-Maverick	0.0%	0.0%	0.0%	0.0%	0.0%	7.7%
Mean tool calls per rollout						
Gemma-3-27B	3.39	1.72	2.19	1.97	3.04	11.21
Sarvam-M	5.11	2.26	2.60	1.48	3.51	9.94
GPT-OSS-120B	1.52	0.66	0.43	0.53	0.45	1.24
Qwen3-235B	4.68	4.22	6.40	4.54	5.91	9.58
Llama-4-Maverick	4.46	2.32	3.94	2.62	3.96	7.84
Median response length (characters)						
Gemma-3-27B	1,004	362	906	1,121	1,067	5,679
Sarvam-M	1,511	658	1,299	642	1,302	3,265
GPT-OSS-120B	**117**	**31**	**33**	**33**	**38**	463
Qwen3-235B	971	1,092	1,375	1,742	1,178	4,653
Llama-4-Maverick	1,655	880	2,079	1,564	1,622	4,850

Table 13: Bold entries mark the pathology of §6: GPT-OSS's two *highest* raw invariance scores sit on its two highest parse-failure rates, and its median response is one short sentence.

A trace consisting of a single `Finish` call scores perfect invariance while demonstrating no agentic behaviour. These are common: Sarvam produces them on 8,879 of 14,400 Belebele rollouts (61.7%), Gemma on 7,215 of 14,268 XQuAD rollouts. Any invariance number that does not exclude them is partly a measure of degeneracy.

Tool-name hallucination is negligible ($\leq$ 0.11% of calls for every model). Across the release, 1,412,535 tool calls were emitted, of which 389 were non-canonical (most commonly `calculate` for `Calc`). The action alphabet is therefore genuinely shared, which is what makes cross-lingual trace comparison well-posed.

L GPT-OSS-120B Failure Taxonomy

This is the paper's clearest single demonstration that an evaluation artifact can be mistaken for a model property, so we document it in full. The question the table below answers is narrow but consequential: when a model's measured accuracy moves by more than an order of magnitude under an intervention that touches neither its weights nor its decoding, which of the two numbers was ever measuring capability?

Per-benchmark few-shot accuracy for GPT-OSS runs Belebele 0.023 $\rightarrow$ 0.399 $\rightarrow$ 0.362; XCOPA 0.026 $\rightarrow$ 0.530 $\rightarrow$ 0.579; XNLI 0.017 $\rightarrow$ 0.513 $\rightarrow$ 0.570; XQuAD 0.007 $\rightarrow$ 0.384 $\rightarrow$ 0.394 (0-shot $\rightarrow$ 2-shot $\rightarrow$ 4-shot). Qwen3, already legible, gains far less: Belebele 0.782 $\rightarrow$ 0.885 $\rightarrow$ 0.892, XCOPA 0.855 $\rightarrow$ 0.919 $\rightarrow$ 0.929, XNLI 0.611 $\rightarrow$ 0.756 $\rightarrow$ 0.750.

Under strict empty exclusion, 89.93% of GPT-OSS's within-language pairs vanish, leaving $n_{\text{within}} = 24$ on XCOPA and 68 on XNLI, and 3 of its 6 cells have intervals crossing zero. Its apparent gap of $+0.0450$ measures the minority of cases where it followed the protocol at

	Adherence (GPT-OSS-120B)			Correctness (gold benchmarks)		
Arm	Empty	Parse fail	Usable pairs	Scorable	Accuracy	Acc. | scorable
0-shot (baseline)	74.4%	59.5%	10.4%	2.1%	0.0174	**0.8127**
2-shot	28.7%	21.7%	53.3%	58.4%	0.4421	0.7574
4-shot	**27.8%**	**19.7%**	**55.6%**	**61.2%**	**0.4539**	**0.7418**

Table 14: Format exemplars repair measurement, not capability. Measured accuracy rises 26× (orange to green) while accuracy among scorable rollouts *falls* slightly: the 0-shot 0.8127 rests on $n = 299$ heavily selected rollouts, the 4-shot 0.7418 on $n = 8{,}567$. The effect saturates at two exemplars, which is itself informative: if the exemplars were teaching the task rather than the format, four would beat two.

Benchmark	Rollouts	No trace	Non-empty	Names a tool	% of non-empty
FLORES-200	20,937	10,692	9,842	9,745	99.0%
XQuAD	14,268	10,767	7,299	6,673	91.4%
XNLI	37,350	33,095	29,700	26,672	89.8%
Belebele	14,400	11,500	8,614	7,248	84.1%
XCOPA	1,100	974	804	772	96.0%
Synthetic	2,300	1,970	1,299	948	73.0%
Total	**90,355**	**68,998**	**57,558**	**52,058**	**90.4%**

Table 15: The final column is the share of *non-empty* failures, and must never be quoted without that qualifier. Over *all* failures the recovery ceiling is 52,058/68,998 = 75.4%; the residual 16.6% are genuinely empty responses. Post-recovery adherence would be 81.3%, up from 23.6%.

all, which is why it is excluded from the headline pooling. §6 shows this is a *measurement* decision, not a capability judgement.

M Temperature Ladder

The ladder exists because two points do not distinguish "the gap grows at T=0" from "the ceiling moves at T=0". Five rungs on Gemma-3-27B and four on Sarvam-M, with model, task set and seed fixed, separate the two: I_{cross} is read *within* a model across temperature, which is the only comparison the single-replicate rungs support. Ceiling-corrected columns exist only where two replicates were generated.

	I_{cross}			Behaviour		Ceiling-corrected		
T	Raw	A: →0.5	B: 0.5→	Len.	Empty	I_{within}	Gap	$\tilde{I}$
Gemma-3-27B								
0.0	0.6635	0.6611	0.6632	3.48	0.04%	0.9110	+0.2482	0.7276
0.3	0.6632	0.6608	0.6631	3.45	0.04%	—	—	—
0.5	0.6608	0.6608	0.6608	3.44	0.04%	0.7957	+0.1344	0.8311
0.7	0.6565	0.6576	0.6597	3.39	0.04%	0.7623	+0.1191	0.8438
1.0	0.6546	0.6576	0.6576	3.38	0.05%	—	—	—
Sarvam-M								
0.0	0.6226	0.6162	0.6025	3.88	1.79%	0.8797	+0.2572	0.7076
0.3	0.6153	0.6099	0.6013	3.90	1.93%	—	—	—
0.5	0.5958	0.5958	0.5958	3.84	2.42%	0.6477	+0.0506	0.9219
0.7	0.5697	0.5772	0.5884	3.77	4.26%	0.5977	+0.0400	0.9330

Table 16: $T = 0.3$ and $T = 1.0$ are single-replicate by design, so no ceiling correction exists there; the flatness claim rests on I_{cross}, used only as a within-model comparison across temperature with model and task set fixed. Sarvam's $T = 1.0$ arm was not generated. **The $\tilde{I}$ column is computed on length-matched I_{cross} whereas the "Raw" column is not**, so dividing the two neighbouring columns does not reproduce it: at T=0.7 Gemma's raw ratio is 0.861 against the length-matched 0.8438 printed here.

N English-Pivot Ablation

The ablation ran in two campaigns under one protocol. Gemma-3-27B and Qwen3-235B were run first (24 cells, 90,400 rollouts) and produced a *post-hoc* head-room account of why the mandate helped one and not the other. Llama-4-Maverick and Sarvam-M were then selected *on that account*, from their measured baseline pivot rate rather than from any outcome, and both predictions were written into the analysis script before the compute was spent (36 cells, 135,600 rollouts, including an in-job baseline arm so that all three arms are matched by construction rather than borrowed). **Sixty cells and 226,000 rollouts in total**, all matched on seed 101, T=0.5, 4096 tokens and 300 tasks, with the same two intervention prompt files used verbatim for all four models so the manipulation is textually identical rather than merely equivalent.

N.1 The pre-registered test, consolidated

Table 17 is the four-model result in one place. The upper two rows of each block are the original campaign, the lower two the pre-registered extension; reading the "Bench" columns downward is the test.

			`Translate` removed			`Translate` mandated		
Model	Head-room	Moved	A	B	Bench	A	B	Bench
Sarvam-M‡	**57.2**	+38.6	+0.0015	+0.0064	1/6	+0.0110	+0.0190	**5/6**
Gemma-3-27B	30.4	+30.0	**−0.0432**	**−0.0348**	4/6	**+0.0635**	**+0.0674**	**5/6**
Llama-4-Maverick	18.7	+17.4	−0.0511	+0.0124	4/6	+0.0242	+0.0154	3/6
Qwen3-235B	15.7	+14.7	−0.0229	+0.0659	4/6	−0.0105	−0.0117	2/6

Table 17: All four ablated models, ordered by head-room; the two lower rows in each block were predicted before generation. Head-room is 100 minus the baseline first-action `Translate` rate; "Moved" is the increase in that rate under the mandate. A and B are the two length-matching directions on I_{cross}; orange marks a pooled estimate that fails the both-directions test, in which case the per-benchmark counts are the honest read. ‡Sarvam's mandate reached only 81.4% first-action compliance against 98.7–99.6% elsewhere, so part of its small magnitude may be weak manipulation rather than a model property; this is the largest single threat to the magnitude reading below.

Direction is predicted; magnitude is not. Ranked by benchmarks where the mandate helps with both directions agreeing, the four models give 5/6, 5/6, 3/6 and 2/6, non-increasing in head-room, on four independent systems, with the two new points fixed in advance. The pooled magnitudes do not follow: Gemma +0.0655, Llama-4 +0.0198, Sarvam +0.0150, Qwen3 −0.0111. Sarvam has the most head-room and *moved furthest* (+38.6 points against Gemma's +30.0), yet gained 4.4× less, and Llama-4 is third on head-room but second on magnitude. We therefore report head-room as governing *whether* the mandate helps, not how much, and offer no prescription.

Llama-4's removal arm is the third instance of the raw-versus-matched inversion, and is the cleanest case in the paper. Raw I_{cross} *rises* 0.6484 → 0.7038 (+0.0554), which read alone would reverse §5 outright; removal shortens its traces by 24% (4.08 → 3.09 actions), and matching to the baseline length distribution flips the sign to −0.0511. That the inversion recurs on a third model, in a campaign designed after the first two, is why every comparison in this paper is length-matched in both directions.

Two pooled estimates fail the both-directions test, for a reason worth stating. Direction B reweights the baseline onto the arm's length distribution, which is unreliable when the arm populates a bin the baseline barely reaches. Llama-4's removal arm puts 1,593 cross-language pairs in the shortest bin where its baseline has only 582, and its mandate arm only 65; Sarvam, by contrast, has at least 10,378 pairs in every bin of every arm. Sarvam's estimates are therefore the best-supported in the set, which matters because Sarvam is the model that supplies the negative result below.

N.2 The removal finding does not generalise, and that is the informative part

Across two models, removal lowered length-matched cross-lingual agreement in both. Across four it does not. **Llama-4 replicates it per benchmark**, 4/6 negative with both directions agreeing, even though its pooled estimate fails the both-directions test. **Sarvam-M contradicts it**: only 1/6 benchmarks is negative and the pooled estimate is +0.0015 / +0.0064, which at one seed per arm should be read as *no effect* rather than as a small positive one. The coherent reading is that the pivot is load-bearing *in proportion to how much a model uses it*: Sarvam opens with `Translate` on 42.8% of non-English rollouts, roughly half of the other three, so there is less pivoting to remove and removal costs it nothing. That is a dose–response refinement of the mechanism rather than a refutation of it, but a flat claim that removal lowers agreement "in models" is not available, and we do not make one.

N.3 Per-benchmark length-matched deltas

Each entry is the pair of length-matched deltas (direction A / direction B) for that benchmark, and a cell counts toward the "Net" row only when the two directions agree in sign.

Benchmark	Gemma removed	Gemma mandated	Qwen3 removed	Qwen3 mandated
FLORES-200	−0.054 / −0.054	+0.037 / +0.076	−0.006 / +0.262 ×	−0.010 / −0.001
XQuAD	+0.009 / +0.007	+0.133 / +0.215	−0.027 / −0.014	+0.088 / +0.151
XNLI	−0.245 / −0.085	+0.021 / +0.022	−0.170 / −0.038	−0.062 / −0.043
Belebele	−0.047 / −0.031	+0.158 / +0.167	−0.122 / −0.072	+0.046 / +0.065
XCOPA	−0.227 / −0.168	+0.016 / +0.014	−0.308 / −0.210	−0.166 / −0.041
Synthetic†	+0.051 / +0.052	−0.022 / −0.017	+0.089 / +0.114	−0.149 / −0.062
Net	**4/6 negative**	**5/6 positive**	**4/6 negative**	2/6 positive
Benchmark	Llama-4 removed	Llama-4 mandated	Sarvam removed	Sarvam mandated
FLORES-200	−0.133 / −0.069	+0.034 / +0.037	−0.067 / −0.050	+0.039 / +0.048
XQuAD	+0.065 / +0.079	+0.031 / +0.098	+0.002 / +0.001	+0.009 / +0.081
XNLI	−0.104 / −0.060	−0.059 / −0.007	+0.052 / +0.066	+0.068 / +0.074
Belebele	−0.024 / +0.085 ×	+0.016 / +0.052	+0.005 / +0.005	+0.001 / +0.015
XCOPA	−0.088 / −0.051	−0.016 / +0.018 ×	+0.091 / +0.112	+0.093 / +0.105
Synthetic†	−0.006 / −0.004	−0.020 / −0.017	+0.017 / +0.017	−0.045 / −0.038
Net	**4/6 negative**	3/6 positive	1/6 negative	**5/6 positive**

Table 18: × marks a cell where the two matching directions disagree in sign and which therefore does not count. Sarvam's per-benchmark agreement is 12/12, covering every cell, both arms and both directions, against Llama-4's 10/12. Two patterns cut across models. **Synthetic inverts under the mandate in all four**, and it is also one of the two benchmarks where removal *raises* agreement in three of them; both are consistent with its far lower chance floor (0.351 against 0.57–0.64; Appendix Q) meaning it measures something different from the natural benchmarks. **XQuAD is the other dissenter for removal**, positive in three of four models. On every other benchmark, removal is negative wherever a model pivots heavily. †**The Qwen3 synthetic cell additionally carries a truncation confound and should be discounted**: mean trace length is 9.53 (baseline), 10.44 (removed) and 3.69 (mandated), with truncation 21.6%, 34.8% and 6.6%. The 20.7–21.1% empty rate there is the pre-existing low-resource adherence collapse and is stable across arms.

N.4 Tool redistribution under the ablation

Share of all tool calls, as `Search` / `Calc` / `Translate` / `Summarize` / `Finish`. Gemma baseline 10.2 / 21.3 / 28.3 / 11.9 / 28.2%; removed, 25.7 / 25.7 / 0.0 / 15.9 / 32.7%; mandated, 9.5 / 15.0 / 30.2 / 18.6 / 26.7%. Qwen3 baseline 5.0 / 10.7 / 50.5 / 20.2 / 13.6%; removed, 15.8 / 15.9 / 0.0 / **54.4** / 13.9%; mandated, 6.3 / 4.6 / 50.1 / 20.3 / 18.7%. Llama-4 baseline 10.2 / 11.3 / 35.6 / 18.8 / 24.1%; removed, 14.8 / 13.8 / 0.0 / **39.8** / 31.7%; mandated, 14.2 / 10.7 / 36.9 / 17.5 / 20.6%. Sarvam baseline 23.5 / 16.1 / 18.0 / 18.0 / 24.2%; removed, **32.2** / 20.2 / 0.0 / 21.7 / 25.9%; mandated, 22.0 / 13.5 / 25.4 / 16.5 / 22.3%. Calls to a tool that was not offered number 6 in ∼500,000 (0.01%), so the alphabet restriction was respected in every model.

The freed budget is reallocated *differently*, and the difference tracks the invariance result. Qwen3 sends 68% of the withdrawn share to `Summarize` (20.2 $\rightarrow$ 54.4) and Llama-4 59% (18.8 $\rightarrow$ 39.8), so both substitute another English-producing operation. Gemma splits, 55% to `Search` and 16% to `Calc`. Sarvam sends only 20% to `Summarize` and 48% to `Search` (23.5 $\rightarrow$ 32.2). In every model the gains reconcile to the withdrawn share to within 0.2 points, so this is a genuine reallocation rather than a counting artifact. The two models that lean hardest on the pivot replace it with the one remaining tool that also emits English; the model where removal costs nothing reaches instead for retrieval.

In the reasoning-language control, `Thought:` text in the task's script, by arm: Gemma 0.09% (unconstrained), 0.06% (think-in-English), 0.79% (think-in-task-language); Sarvam 0.16% / 0.04% / 0.08%. Mean ASCII fraction of `Thought:` text on non-Latin-script arms is 0.986–0.993 in every condition. The think-in-English arm's length-matched effect is $+0.011$ (Gemma) and $+0.014$ (Sarvam), with both matching directions agreeing.

O Self-Consistency Voting

Voting is the most obvious repair a practitioner would reach for, so it deserves a careful report rather than silence. Because majority voting reduces variance, it must raise same-language agreement as well as cross-language agreement, and C4 says a gain read without its ceiling is uninterpretable. Testing $\tilde{I}$ under voting needs $2k$ replicates so that *both* sides of the ratio are themselves voted. We therefore ran ten replicates, and the question is now settled.

Voting is a variance reducer, not a retention improver. With vote A (seeds 701–705) and vote B (seeds 706–710) at T=0.7 on 300 tasks, $\tilde{I}$ at k=5 is formable for the first time. For Gemma-3-27B it falls from 0.8481 $[0.8438, 0.8527]$ at k=1 to 0.8323 $[0.8274, 0.8372]$ at k=5; for Sarvam-M from 0.9296 $[0.9238, 0.9353]$ to 0.9109 $[0.9062, 0.9159]$. Both intervals are disjoint. The mechanism is visible in the components: voting raises I_{within} by $+0.078$ and $+0.071$ against I_{cross} by $+0.052$ and $+0.053$: a five-way vote makes a model agree with *itself* more than it makes it agree *across languages*. Two qualifications travel with this. The effect is **small** (-0.016, -0.019), and the intervals are disjoint partly because task-level bootstraps within a model are narrow by construction. And it is measured at T=0.7 only, since voting is undefined at T=0 where all replicates coincide, so it cannot be checked at the temperature where the retention regularity is defined. Both readings are true: raw cross-language agreement does rise $+0.053$, and a practitioner whose objective is raw agreement rather than retention relative to a model's own ceiling will obtain it.

The table below reports the *five*-replicate run (seeds 701–705, T=0.7, majority vote over tool sequences), which measures raw I_{cross} under voting. The ceiling-corrected result above uses the separate *ten*-replicate run, since forming $\tilde{I}$ needs both sides of the ratio voted; the two are different experiments and their numbers should not be compared directly.

P Correctness

Correctness enters this paper for one reason: to test whether trace divergence is merely cosmetic. It is not, and the detail below also shows why invariance cannot be substituted for accuracy as a cheap evaluation signal.

Gold is available for XQuAD (extractive, per-language), XNLI (label), Belebele (MCQ) and XCOPA (MCQ). Scoring is deliberately lenient (first standalone digit for MCQ, first NLI label mentioned, normalised containment for extractive QA), which biases *against* finding a spurious accuracy–invariance relationship.

The accuracy–invariance relation reverses in five cells: Gemma·XNLI -0.477, Llama-4·Belebele -0.669, Llama-4·XNLI -0.376, Qwen3·Belebele -0.299, Qwen3·XNLI -0.129. Median across the 20 adherent cells is $+0.513$.

Model	Benchmark	k=1	k=3	k=5	$\Delta(k_5-k_1)$	Unanimous
Gemma-3-27B	FLORES-200	0.6319	0.6718	0.7283	+0.0964	13.8%
	XQuAD	0.7276	0.7502	0.7703	+0.0427	44.5%
	XNLI	0.7411	0.7522	0.7585	+0.0174	47.3%
	Belebele	0.6885	0.7134	0.7298	+0.0414	27.2%
	XCOPA	0.7416	0.7468	0.7515	+0.0099	46.9%
	Synthetic	0.4688	0.4693	0.4738	+0.0050	2.0%
	pooled	**0.6428**	**0.6645**	**0.6898**	**+0.0470**	**27.8%**
Sarvam-M	Belebele	0.6914	0.7423	0.7957	+0.1043	16.9%
	XQuAD	0.6534	0.6890	0.7418	+0.0884	9.2%
	FLORES-200	0.5361	0.5463	0.5862	+0.0501	0.4%
	XNLI	0.5347	0.5434	0.5714	+0.0367	1.3%
	XCOPA	0.5364	0.5369	0.5651	+0.0288	0.5%
	Synthetic	0.3891	0.3846	0.3801	−0.0090	0.1%
	pooled	**0.5568**	**0.5706**	**0.6040**	**+0.0472**	**5.1%**

Table 19: The gain is largest at *intermediate* replicate unanimity, not at either extreme: mean Δ is +0.096 in the 5–20% unanimity band against +0.022 below 5% and +0.028 above 20%. Where replicates already agree there is nothing to vote away, and where they disagree completely there is no majority to find, so the relationship is an inverted U rather than a monotone one; the pooled linear correlation is correspondingly near zero ($r = -0.08$). Sarvam never exceeds 16.9% unanimity and therefore sits entirely on the rising limb. **Pooled rows are computed over pairs, not as the mean of the cells above them**, so they need not equal that mean (Gemma: +0.0470 pooled against +0.0355 macro-averaged), the same distinction noted for Table 2 in Appendix I. Because voting also raises I_{within} and $k = 5$ cannot supply two independent votes, these gains are **not** evidence that voting improves cross-lingual agreement (§6).

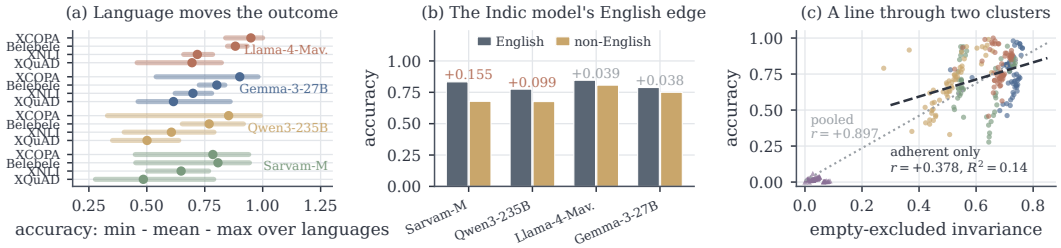

Figure 9: Correctness moves with language, but invariance does not predict it. (a) Within a single model and benchmark, accuracy spans up to 0.66 across languages. (b) Every model has an English advantage, and the Indic-specialised model has the largest. (c) The pooled correlation ($r = +0.897$) is an artifact of two clusters; excluding GPT-OSS it falls to $r = +0.378$.

Q The Chance Floor

S is a normalised matching-block similarity over a five-symbol alphabet, so two traces that answer *different* tasks still agree substantially by construction. Under the standard attenuation model $S_{\text{obs}} = c + (1-c)\, S_{\text{true}}$, the difference $\Delta_{\text{obs}} = (1-c)\, \Delta_{\text{true}}$ rescales cleanly and its ordering is invariant to c, but the ratio $\tilde{I}_{\text{obs}} = (c + (1-c) I_{\text{cross}}) / (c + (1-c) I_{\text{within}})$ is biased toward 1 and differentially so across models. The estimand we argue for is therefore strictly more exposed to an uncorrected floor than the quantity it replaces.

Within each (model, benchmark, condition) cell we permute the task → trace assignment *within* a language arm and recompute the ordinary cross-language S. This is the exact null "what does S score when two traces answer different tasks?", and it preserves the language, the model, the length distribution and the empty-trace pattern of the arm while destroying only the task correspondence. Everything else is held identical to the main estimator: the same S, the same strict empty exclusion, the same cross-seed pairing. Because c is length-dependent it is estimated *within* length bin and reweighted to the observed cross-language bin distribution, exactly as the length-matched estimator does. We use 200 permutations and a fixed seed, giving 1.07M–12.6M null pairs per cell over **66 cells**.

Model	Benchmark	Mean	Min	Max	Spread
Llama-4-Maverick	XCOPA	0.949	0.840	1.000	0.160
	Belebele	0.882	0.848	0.933	0.086
	XNLI	0.718	0.659	0.784	0.125
	XQuAD	0.695	0.458	0.822	0.364
Gemma-3-27B	XCOPA	0.901	0.540	0.980	**0.440**
	Belebele	0.802	0.726	0.837	0.111
	XNLI	0.699	0.622	0.780	0.157
	XQuAD	0.615	0.462	0.860	0.399
Qwen3-235B	XCOPA	0.853	0.330	0.990	**0.660**
	Belebele	0.769	0.653	0.917	0.263
	XNLI	0.606	0.402	0.790	0.388
	XQuAD	0.501	0.352	0.638	0.287
Sarvam-M	Belebele	0.807	0.452	0.943	**0.491**
	XCOPA	0.785	0.450	0.940	0.490
	XNLI	0.648	0.503	0.767	0.263
	XQuAD	0.485	0.278	0.789	**0.511**

Table 20: Per-language accuracy range within each model×benchmark cell. GPT-OSS is omitted here: its accuracy is adherence-limited (§6) and reported separately.

The floor is high, and benchmark-specific. Mean $c = 0.561$ at T=0 (per-cell range 0.266–0.753) and 0.559 at T=0.5. By benchmark at T=0: Belebele 0.644, XCOPA 0.626, XNLI 0.591, FLORES-200 0.585, XQuAD 0.570, synthetic 0.351: the ordering tracks trace length, since the synthetic benchmark has by far the longest traces and correspondingly the lowest collision rate. Across the three P6-A2 models the range widens to 0.314–0.947, and **a single model owns both extremes**: Qwen3-8B scores 0.314 on synthetic and 0.947 on Belebele. Within one model, across six benchmarks, with decoder and task cap fixed, the floor swings 3×. This is the sharpest form of the statement that no benchmark-level comparison on the raw metric is interpretable.

Writing $\kappa = (S - c)/(1 - c)$, chance-corrected retention is 15–18% rather than 71–73% chance-inclusive: of the agreement a model achieves above chance when asked the same question twice in one language, roughly one sixth survives a change of language. The frontier band is 3.0 absolute points before correction and 3.0 after. **In relative terms it widens, from** 4.1% **to** 18.8%, because the correction shrinks the level roughly fivefold while leaving the between-model differences nearly intact; we report both metrics rather than the absolute one alone. The measured floor is *model-specific* (Qwen3-235B 0.450 against Gemma-3-27B 0.623, since Qwen3's traces are longest), and those differences very nearly cancel the differences in raw $\tilde{I}$. A sensitivity analysis that assumes one uniform c for every model therefore misestimates the risk: it predicts the band breaks near $c \approx 0.3$, whereas at the measured $c \approx 0.56$ the band is unchanged.

No cell has $I_{\text{cross}} \leq c$, so no ratio is sign-flipped or undefined; the tightest margin in the set is Qwen3-8B on Belebele at $+0.0024$. Corrected values span 0.023–0.689 with no divergence. The correction divides by $1 - c$, so cells with a very high floor are numerically delicate: at $c = 0.9473$ the divisor is 0.0527, and a 0.005 error in c would move that cell's corrected value from $+0.046$ to -0.055. **No conclusion rests on such a cell**: Qwen3-8B falls outside the corrected band on its pooled value across all six of its cells, not on Belebele alone.

R Scale and Vendor Generalisation

Three systems were chosen so that a break in the frontier regularity would be interpretable rather than merely observed: **Gemma-3-4B** holds family and training recipe fixed against Gemma-3-27B at 6.75× scale difference; **Qwen3-8B** does the same for Qwen at 29×; **Aya-Expanse-8B** is a fourth vendor with explicitly multilingual post-training, the rung most likely to break a training-recipe account. Each was run for two replicates at T=0 on the 300-task subset (precisely the grid cell that produced the 71–73% figure), for 36 cells and 135,600 rollouts.

Pooled over pairs, $\tilde{I}$ is Qwen3-8B 0.8016 $[0.7932, 0.8092]$, Gemma-3-4B 0.6259 $[0.6192, 0.6323]$ and Aya-Expanse-8B 0.4738 $[0.4658, 0.4811]$, against the frontier band $[0.708, 0.733]$. **At least two of the three fall outside the band on every treatment we tried**: three of three under pooled-raw, two of three under macro-of-ratios-raw, and two of three chance-corrected. The

minimum across treatments is two, which is the strongest form of the claim that is true under all of them.

Three aggregations of $\tilde{I}$ are defensible and we state which we use. Pooled over pairs gives the band $[0.708, 0.733]$ (width 2.6 pts); the ratio of macro-averaged cell means gives $[0.713, 0.746]$ (3.2 pts); the macro average of per-cell ratios gives $[0.713, 0.742]$ (3.0 pts). The *width* is stable at 2.6–3.2 points under all three, so the regularity does not depend on the choice, but Gemma-3-27B and Qwen3-235B swap ends of the band between pooled and macro, so no claim about which model retains most is made anywhere in the paper. The main text uses the pooled aggregation throughout.

A cell in which most rollouts are empty measures adherence, not retention, so we re-pool over cells that actually produced traces. For the two clean models the gate barely moves the estimate and never toward the band: Gemma-3-4B 0.6259 $\rightarrow$ 0.6307 at $\leq$ 5% empty (5/6 cells retained), Qwen3-8B 0.8016 $\rightarrow$ 0.8094 (5/6). For Aya the gate *lowers* the estimate to 0.4460 and retains only 3 of 6 cells at $\leq$ 20% empty.

Aya-Expanse-8B is an adherence failure rather than a retention measurement. Its task-weighted empty-trace rate is 31.2%: FLORES-200 86.3%, synthetic 82.1%, XNLI 20.4%, XQuAD 14.0%, Belebele 11.9%, XCOPA 9.5%. On its two worst benchmarks the model emits no parseable trace in more than four rollouts in five. We treat it exactly as GPT-OSS-120B is treated in Appendix L: reported, excluded from the regularity, and used as evidence for the adherence story rather than the retention story. The fourth-vendor question therefore remains open.

Length does not explain the break. Mean trace length at T=0 is Qwen3-8B 1.99, Gemma-3-27B 3.48, Sarvam-M 3.88, Llama-4 4.07, Gemma-3-4B 4.44, Qwen3-235B 5.70. Regressing $\tilde{I}$ on mean length across the six protocol-adherent models gives $\tilde{I} = 0.817 - 0.0254\,\ell$ with $r = -0.547$, $R^2 = 0.30$; per cell across 34 clean cells, $R^2 = 0.15$. The structural refutation is Gemma-3-4B: at 4.44 its mean trace length sits *inside* the frontier range $[3.48, 5.70]$, between Llama-4 and Qwen3-235B, and it still misses the band by 8.2 points. A model at frontier trace length with non-frontier retention is an observation the length hypothesis cannot absorb.

Variance collapses with scale. Across the four frontier models at 17–235B the spread is 2.6 points (SD 0.011); across the two clean models at 4–8B it is 17.6 points (SD 0.088): a 6.8$\times$ difference in spread and an 8.0$\times$ difference in SD. We offer this as a hypothesis at $n = 2$ below 10B. Note also that both clean new models fall inside the per-cell range $[0.61, 0.82]$ already reported for the frontier four: they break the *per-model* band while sitting within the *per-cell* variation the study documents. This is a raw-scale observation and is not comparable to the chance-corrected values above.

S Limitations

What follows is what the study leaves open. Four questions that would otherwise belong here are settled by experiment instead, and appear as results: trace-length causality (Appendix J.1), the chance floor (Appendix Q), the head-room account (Appendix N) and ceiling-corrected voting (Appendix O).

Calls are parsed and compared but never dispatched, and no observation is fed back into the loop. We therefore measure the *induced* tool-use policy under a controlled scaffold, not grounded execution. Whether the same divergence appears when tool outputs change the state is untested and is the most important extension.

The frontier band rests on four systems, and the regime below it on two. $\tilde{I} \in [0.708, 0.733]$ across four frontier models is a striking regularity, but $n = 4$, and the reported intervals are task-level *within* each model, so they establish that each $\tilde{I}$ is precisely estimated rather than that four systems suffice to establish a constant. The per-cell decomposition ($n = 24$, model identity $\eta^2 = 5.7\%$) is the stronger form of the evidence and is what §4 relies on. Extending to eight models locates a boundary but does not close the question: the claim that retention is variable below roughly 10B rests on **two** clean models there (Gemma-3-4B and Qwen3-8B), which is a hypothesis, not a result. A frontier model outside the band, or a

sub-10B model inside it under chance correction, would bound the regime differently. The uniformity is also specific to greedy decoding: at $T{=}0.5$ and $T{=}0.7$ the ratio inflates toward 1 and the models spread apart, so the figure must always be quoted with its temperature.

The fourth-vendor question is open, because the fourth vendor did not adhere. Aya-Expanse-8B was chosen precisely as the rung most likely to break a training-recipe account of the regularity, and it emits no parseable trace on 31.2% of its rollouts, more than four in five on its two worst benchmarks. We therefore report it as an adherence result and exclude it from the retention comparison, exactly as GPT-OSS-120B is treated. The consequence is that all six retention datapoints come from three vendors, and *whether an explicitly multilingual post-training recipe changes retention remains untested*. This is the single most consequential gap in the boundary analysis, and it is a gap we created by our own eligibility rule rather than one the data settled.

Trace length is causal but its sign is not universal. The three-level manipulation establishes that $\tilde{I}$ moves 6–7 points under a $3\times$ change in length, with disjoint intervals, further than the entire across-model band. But the two models tested *disagree on the direction*: Qwen3 is monotone decreasing, Gemma is not, and its anomaly is the shortest arm. A third model is needed before anything general can be said about which way length pushes retention, and a four- or five-level ladder before the non-monotonicity can be distinguished from noise. The instruction is also a soft prompt constraint rather than a decoding limit, so the arms overlap in the tails.

The think-in-task-language arm was refused by both models ($< 1\%$ compliance), so the designed contrast never existed. This must be reported as a *failed manipulation*, not as evidence that reasoning language does not matter. That hypothesis remains untested, and the refusal itself is the finding.

The pivot ablation settles less than four models might suggest. The *removal* direction does not generalise. It lowers length-matched agreement on $4/6$ benchmarks in three models and on $1/6$ in the fourth, which we read as dose–response, with the pivot load-bearing in proportion to use, but that reading is inferred from four points rather than tested. The *mandate* direction was pre-registered and its *ordering* confirmed, yet its *magnitude* is monotone in neither head-room nor movement, so we offer no prescription. Three specific weaknesses travel with it. Sarvam-M's mandate reached only 81.4% first-action compliance against 98.7–99.6% elsewhere, so part of its small magnitude may be weak manipulation. Head-room is measured on *first-action* `Translate` alone and does not capture pivoting later in a trace or inside reasoning text, and the reasoning-language result above shows that internal pivoting is prompt-resistant. And each arm has one seed and one replicate, so pooled differences below roughly ± 0.01, which includes Sarvam's removal result, should be read as "no effect" rather than as a small positive one; no $\tilde{I}$ or gap is formable for these arms, only I_{cross}.

Voting is measured at one temperature only, because it is undefined at $T{=}0$, where all replicates coincide, so the ten-replicate experiment runs at $T{=}0.7$. The finding that voting costs 1.6–1.9 points of $\tilde{I}$ therefore *cannot be checked at the temperature where the retention regularity is defined*, and the intervals are disjoint partly because task-level bootstraps within a model are narrow by construction. The effect is small in absolute terms, and a practitioner whose objective is raw cross-language agreement rather than retention relative to a model's own ceiling does obtain it ($+0.05$).

Finally, S is a single sequence-similarity metric. The chance floor it lacks is now measured rather than assumed (Appendix Q), but the choice of S itself is not varied: whether an alternative trace-similarity family, such as edit distance under a tool-cost matrix or a set-based measure insensitive to order, would preserve the frontier band is untested. Absolute values of I_{cross} should not be compared across metric definitions.

Gold exists for four of six benchmarks, and the two excluded (FLORES-200, synthetic) are the longest-trace ones, so the correctness picture comes from the shorter half of the suite. Scoring is deliberately lenient, which makes absolute accuracies upper bounds; because the same scorer applies everywhere, cross-language comparisons are unaffected. Invariance and

accuracy are measured on the same rollouts but are not interchangeable: among adherent models the association is $r \approx +0.38$ and it reverses in a quarter of cells.

Sarvam-M has no T=1.0 rung on the temperature ladder, so the flatness claim of §3 rests on the full five-point range in one model and four points in the other; its think-natively arm covers one benchmark rather than six. Neither affects a headline number, and both are visible in Appendices M and N rather than absorbed into an average.

T Prompts

The baseline scaffold is fixed across all languages and models:

```
You are an agent that solves tasks using tools.

Available tools:
- Search(query: str) - Search for information
- Calc(expression: str) - Calculate mathematical expressions
- Translate(text: str) - Translate text to English
- Summarize(text: str) - Summarize text
- Finish(answer: str) - Provide final answer

Rules:
- You must use tools for reasoning.
- Do NOT give direct answers without tools.
- Always follow this format:

Thought: ...
Action: ToolName(arguments)

Repeat until final answer using Finish().
Task: {task}
```

Interventions modify exactly one element. **Tool removed**: `Translate` is deleted from the tool list. **Tool mandated**: a rule is added requiring `Translate` as the first action when the task is not in English. **Reason in English** / **reason natively**: a rule is added requiring every `Thought:` line to be in English, or in the task's language. **Few-shot**: two or four complete worked `Thought`/`Action` examples are appended before the task. With no intervention set, the scaffold hashes byte-identically to the baseline used in every run, which is verified programmatically.

U Compute

All generation ran on nodes of 8× NVIDIA B200 (192 GB) under vLLM 0.18 (Kwon et al., 2023), with tensor/data-parallel layouts chosen per model (1×8 for Gemma, Sarvam and GPT-OSS; 8×1 for Qwen3 and Llama-4). The main sweep required 2h30m to 4h13m per model, and the full campaign is approximately 120 GPU-hours of B200 time. Every condition writes to an isolated, labelled results tree, and 53 automated checks assert that no labelled run can be read as a baseline cell or resume from one. Analysis is CPU-only and deterministic given a fixed bootstrap seed.

V Provenance

An analysis that reads the wrong results directory produces a plausible number from incomplete data, silently. Two properties therefore have to hold, and we check both rather than assume them: every arm of the campaign must be exactly complete, and every reference to an arm must resolve to exactly one directory. Table 21 reports the outcome.

Benchmarks are identified by content rather than by directory name. Run tags are length-capped, so a long model identifier can push the benchmark suffix off the end of a cell directory name, and for one model all five adapted-benchmark cells were named identically apart from their timestamp. A resolver keyed on the name silently drops or collides them, so we key instead on each benchmark's *language count*, which is unique across the suite (FLORES-200 21, XQuAD 12, XNLI 15, Belebele 16, XCOPA 11, synthetic 23) and survives any renaming. Under this resolver all main-sweep results reproduce exactly ($\Delta = 0.0000$).

Check	Outcome
Arms present against planned	**44 / 44** across 21 conditions
Arms holding exactly 6 cells	**44 / 44**
Arms holding exactly 22,600 rollouts	**44 / 44**
Total rollouts in the audited campaign	**994,400**
Arms resolving to more than one directory	**0 of 44**
Cells with $I_{\text{cross}} \leq c$ (chance floor)	**0 of 66**

Table 21: Every arm is exactly complete and every reference resolves unambiguously. An "arm" is one (condition, model) unit; 22,600 rollouts is 300 tasks times the language count of each benchmark, plus the two benchmarks with fewer than 300 available tasks run to completion.

Incomplete generations are excluded structurally rather than by filtering: they are moved out of the analysis tree before any script runs, and the scripts read only the active tree. The resolver additionally raises on a repeated benchmark within a directory, so a duplicate cannot be silently averaged even if one were left in place.