

Sample-Matched Comparisons Overstate Test-Time Compute Gains

Srinivas Harish

New York University

srinivas.harish@nyu.edu

Abstract

Inference-time methods buy accuracy with extra compute: sample several reasoning traces and combine them. They are compared at a matched number of samples, and that protocol hides two things. Samples do not cost equal tokens, so matched samples is not matched budget; and it never asks whether a cheaper source of diversity already reaches the same ceiling. We measure both on cross-lingual aggregation, which votes across traces generated in different languages, using 32,992 traces from Qwen3-8B across 11 languages on two item-parallel corpora. At a matched trace budget on MGSM, one trace from each of the eight strongest languages buys no ceiling and no vote that resampling English eight times does not: the per-item oracle is 0.992 for samples against 0.989 for languages, the vote 0.966 against 0.956, and the language ballot costs $1.09\times$ the tokens. Over the full pool the eighty traces from the other ten languages solve 0 of 250 items that English’s own eight do not. On MMLU-ProX-Lite, far from saturation, the languages do hold coverage English never reaches (0.895 against 0.685 over the full 88-trace pool against English’s eight, 26 items) — and at a matched budget of eight traces the language ballot there is still the weaker vote (0.453 against 0.487). What blocks the conversion is not error correlation, which is low: an independence model built from the measured marginals over-predicts the vote by 5.4 and 48.0 points, because the independence sits on items every language already answers and vanishes on the items that decide the score, where wrong answers concentrate on a single value (0.464 of erring languages). Three selectors computable from the released traces without gold access all fail to beat the best single language. We release both pools with per-trace token counts so any aggregation or routing rule can be scored against these ceilings by re-scoring rather than generating.

1 Introduction

Spending more compute at inference time is the standard way to improve reasoning: sample several traces and return the answer most of them agree on (Wang et al., 2023). Methods in this family differ in where the diversity between traces comes from — resampling at temperature, varying the prompt, or, in the cross-lingual variants we study here, generating in different languages and voting across them (Yu et al., 2025; Elhady et al., 2026). They are almost always compared at a matched number of sampled traces.

Two things that protocol cannot see. First, a trace is not a unit of compute: token cost per trace varies with the language, so a matched-sample comparison quietly hands a subsidy to whichever policy draws the more expensive traces. Second, and less obviously, it never establishes that the expensive diversity is *necessary*. A vote can only return an answer some trace produced, so asking how often *any* trace in a set is right gives a ceiling no rule over that set can exceed — and if a cheaper source of diversity reaches the same ceiling at the same budget, the expensive one buys nothing however the votes are counted. Neither quantity is reported in this literature.

We measure both, and the second is where cross-lingual aggregation fails. At a matched trace budget on MGSM, one trace from each of the eight strongest languages reaches a ceiling of 0.989 against 0.992 for eight temperature samples of English alone, and a vote of 0.956 against 0.966, at $1.09\times$ the tokens; over the full pool the eighty traces from the other ten languages solve 0 items that English’s own eight do not. The multilingual apparatus is not buying coverage on this corpus — resampling one language already has it.

That is a property of the corpus, not a law. On MMLU-ProX-Lite, where the model is far from saturation, the languages do hold coverage English

never reaches: 0.895 of items against 0.685 over the full pool, with 26 items solved only outside English. Diversity across languages is worth buying where the model is weak, and there the question becomes whether a vote can convert it. It cannot — at a matched budget of eight traces the language ballot returns 0.453 against 0.487 for monolingual sampling, *below* the cheaper policy despite the higher ceiling.

A candidate explanation is that the languages fail on the same items, so that additional languages carry no independent evidence. Averaged over items this is false: conditioned on one of two languages being wrong, both are wrong on only 0.215 of items, and when both are wrong they agree on the wrong answer only 0.224 of the time. Both statistics say a vote should work. It does not, and the reason is that neither statistic is uniform across items. A homogeneous independence model built from the measured per-language marginals predicts 1.000 on MGSM against 0.946 observed, and 0.820 against 0.340 on MMLU-ProX-Lite. The independence that would justify voting is concentrated on items the model already answers correctly; on the items that decide the score, many languages fail simultaneously and their wrong answers concentrate, with the single modal wrong answer held by 0.464 of the wrong languages.

Because the obstacle is dispersion rather than dependence, the constructive consequence is a routing problem rather than a bigger ballot, and we show that problem is small: a perfect selector over the two strongest languages reaches 0.980 on MGSM, capturing 80% of the headroom that all 11 languages jointly hold, and three languages capture 100%.

Our contributions:

- **A budget-matched comparison of two sources of diversity.** At a matched trace budget, and reported with token cost so the matched-token reading is available too, temperature resampling within one language and one-trace-per-language ballots are indistinguishable on the corpus where the model is strong, and the language ballot is the more expensive of the two.
- **A ceiling, and the measurement of how little voting reaches it.** The per-item oracle bounds every rule that selects or votes among the same traces. We report it for both diversity sources on both corpora, and show that

where the extra coverage exists the vote does not convert it — an independence model built from the measured marginals over-predicts by 5.4 and 48.0 points, and we localise the discrepancy to item-level heterogeneity and to concentration of wrong answers on hard items.

- **A floor, and a re-scorable artifact.** Three selectors computable without gold access, none of which beats the best single language, together with both generation pools and per-trace token counts, so a new rule can be scored against these ceilings by re-scoring rather than by generating.

2 Related Work

Sampling and aggregation. Self-consistency (Wang et al., 2023) samples multiple chains of thought and returns the majority answer. Its behaviour is governed by how often an individual sample is correct and how correlated the samples’ errors are, and every method that spends inference compute on diversity — temperature resampling, prompt variation, cross-lingual generation — is a choice of where that diversity comes from. Self-consistency in one language is therefore not only a baseline here but the *cheapest available source of the same commodity*, and it is the arm against which we price the others. It is also the arm this literature omits: comparisons are reported against single-sample monolingual accuracy or against another multilingual policy, not against monolingual sampling at the same budget.

Multilingual reasoning. Multilingual chain-of-thought reasoning and the MGSM benchmark are due to Shi et al. (2023). A body of work studies which language a model should reason in, and reports that models often reason in English regardless of the input language. We hold the reasoning language fixed to the item language and study aggregation across those languages, which is a different axis.

Cross-lingual aggregation. Yu et al. (2025) report +2.8 points over English self-consistency on MGSM by integrating eleven languages by majority vote. Elhady et al. (2026) enforce cross-lingual self-consistency as a training signal rather than an inference-time rule. Rajaei et al. (2025) learn a reranker over candidates produced in several languages, which is a selector in our terms rather than

a vote. [Qin et al. \(2023\)](#) translate into a pivot language before reasoning. [Ranaldi et al. \(2024\)](#) take a complementary route, having the model emit executable programs as language-aligned intermediate steps and ensembling reasoning paths across languages; structured intermediates of that kind attack the answer-level dispersion we measure in §5.6 at its source, by making evidence from different languages comparable before it is aggregated rather than after. These works compare policies at a fixed number of sampled traces.

Two properties of that comparison are not reported in this literature and are measured here. First, a trace is not a fixed unit of compute: token cost per trace differs by language, so equal-trace comparisons are not equal-budget comparisons, and we quantify the resulting subsidy. Second, none of these works reports the ceiling — the accuracy attainable if the rule always selected a language that happened to be right — so the headroom their rules leave unused is unknown. Our contribution is orthogonal to any particular rule: it supplies the bound against which all of them should be read.

Cost accounting. Tokenizer fertility varies across scripts and is charged to users directly, and multilingual cost disparity has been documented as an equity concern. We connect it to aggregation: a policy that samples expensive languages is not merely more costly but is being compared unfairly when the protocol counts traces.

3 The Aggregation Ceiling

Let $\mathcal{L}$ be a set of languages, and let items be answered in a space where answers from different languages are directly comparable — a number, or an option letter. Comparability is what makes cross-lingual voting well defined at all, and it is why we use corpora with numeric or multiple-choice answers. For item x and language ℓ , let $\hat{y}_\ell(x)$ be the answer language ℓ returns after aggregating its own k traces, and $y^*(x)$ the gold answer.

Oracle. The per-item oracle

$$\text{Or}_{\mathcal{L}}(x) = \mathbb{K}[\exists \ell \in \mathcal{L} : \hat{y}_\ell(x) = y^*(x)] \quad (1)$$

is 1 when at least one language is correct. Its mean over items upper-bounds every rule that selects or votes among the same per-language answers, because such a rule cannot output an answer no language produced. Call this the *language-level* ceiling; the same construction over individual traces

gives the *trace-level* ceiling, which is what a budget of k traces bounds (§5.1) and is the higher of the two. Restricting $\mathcal{L}$ to a subset gives the ceiling available to a rule that may only consult that subset, which is how we measure the size of the routing problem in §5.7.

Error agreement. For languages a, b we report $\Pr[\text{both wrong} \mid \text{either wrong}]$. A value near 1 means the languages fail on the same items, so a vote has nothing independent to aggregate and the oracle is close to the best single language. A value near 0 means failures are disjoint and independent evidence exists. We prefer this conditional to raw agreement because raw agreement is dominated by the items every language answers correctly.

Token efficiency. With a_ℓ the per-trace accuracy and c_ℓ the mean generated tokens per trace, $e_\ell = a_\ell/c_\ell$ is accuracy returned per generated token. Efficiency, not accuracy, is what a language costs a fixed inference budget.

Two matchings. Under *sample matching* each policy draws k traces; this is the protocol of prior work. Under *token matching* each policy spends the same number of generated tokens. The two coincide only if languages share c_ℓ , and their difference is the subsidy sample matching grants a multilingual policy.

4 Experimental Setup

Model. Qwen3-8B, bfloat16, temperature 0.7, top- p 0.95, in non-thinking chain-of-thought mode, served with vLLM on eight H200 GPUs. Each prompt is the item in its own language followed by a fixed English instruction to reason step by step and emit the final answer on the last line; the instruction is identical across languages so that any per-language difference is a property of the item and the model, not of the prompt.

Corpora. MGSM ([Shi et al., 2023](#)), 250 problems per language with free-form numeric answers, and MMLU-ProX-Lite ([Xuan et al., 2025](#)), ten-option multiple choice, which is far from saturation for this model. Both are item-parallel across languages: an item is the same item in every language, which is what makes the oracle and the agreement statistics well defined. We use the 11 languages common to both: English, Spanish, French, German, Russian, Chinese, Japanese, Thai, Swahili, Bengali, Telugu.

Pools. For every (item, language) pair we draw $k = 8$ traces and record each answer and its generated token count, giving 32,992 traces (22,000 on MGSM, 10,992 on MMLU-ProX-Lite). Every number in this paper is a re-score of that fixed pool: all policies see identical generations and differ only in which traces they consume. Budgets larger than k are simulated by resampling from the pool with replacement, averaged over repetitions.

Scope of each measurement. Per-language statistics and the oracle are computed on both corpora. Policy comparisons require every language to be present for every item; that condition holds for all 250 MGSM items and all 11 languages. The budget-matched comparison of §5.1 needs only the k languages it draws, so it is paired on both corpora and reports its own n ; the eleven-language sweep needs every language at once, so on MMLU-ProX-Lite it is reported for completeness only (Appendix F). Language rankings used by any policy are fit on even item ids and evaluated on odd ids; no ranking is fit on evaluation data.

5 Results

5.1 Two sources of diversity at a matched budget

Table 2 and Figure 1 hold the number of traces fixed and vary only where the diversity between them comes from. On MGSM at $k=8$, eight temperature samples of English reach a ceiling of 0.992 and vote 0.966; one trace from each of the eight strongest languages reaches 0.989 and votes 0.956. The paired differences are 0.003 [0.001, 0.007] on the ceiling and 0.010 [−0.009, 0.029] on the vote, and the language ballot costs $1.09\times$ the tokens. Neither difference favours the languages and neither does the cost. (Three token ratios appear below and are not one quantity: $1.09\times$ here, $1.10\times$ for the fixed- k round-robin of §5.3, and $13.1\times$ for one language to eleven under coverage.)

The full pool makes the same point exactly rather than in expectation. Taking every trace available, English’s own eight samples solve 0.992 of MGSM items and *all* 22,000 traces across 11 languages solve 0.992 — the eighty traces from the other ten languages add 0 items. Whatever the other languages know on this corpus, English already reaches it by being asked again.

MMLU-ProX-Lite behaves differently, and the difference is difficulty. There English’s eight samples solve 0.685 of items and the full pool 0.895,

so 26 items are solved only outside English. At a matched budget the two sources are not separable at this sample size (Δ ceiling -0.018 [−0.119, 0.085], $n=124$), but the coverage is real and it is larger than anything a matched-budget experiment can show, because it appears only once enough traces are bought to reach it. The vote, however, moves the other way: 0.453 for the eight-language ballot against 0.487 for monolingual sampling at the same budget. A policy with the higher ceiling returns the lower accuracy.

5.2 The unmatched comparison, and why we do not rely on it

The *language-level* ceiling over 11 languages is 0.984 on MGSM against 0.944 for a single English sample, and 0.672 against 0.379 on MMLU-ProX-Lite. It sits below the trace-level 0.992 of §5.1 because each language is first collapsed to one answer. Stated that way the premise of cross-lingual aggregation looks secure, and this is how the comparison is usually made; it is also not budget-matched, the left-hand number consuming 11 languages’ worth of traces and the right-hand one a single trace. What a ceiling is good for is the matched comparison instead.

5.3 Ballot size does not rescue it

Appendix B adds languages to the vote in order of validation accuracy under two protocols, and neither raises accuracy. The first is the protocol of prior work, a fixed budget of k traces drawn round-robin; when k is smaller than the ballot it cannot reach every language, so at $k=8$ the nominal eleven-language vote is a vote over the 8 highest-ranked languages, and the three never sampled are the weakest and three of the four most expensive. The second draws one trace per language, so every language votes: accuracy moves from 0.929 at one language to 0.944 at eleven while cost rises $13.1\times$. Reproducing the prior comparison directly, cross-lingual aggregation at $k = 8$ scores -0.7 points against English-only at the same k while spending $1.10\times$ as many tokens, and holding tokens fixed instead moves it to -1.5 . The standard error on a cell is 1.4 points at $n = 250$, so what the data supports is that cross-lingual voting is not better than English-only sampling here, and that the sample-matched protocol reports it while granting it more compute.

Language	MGSM			MMLU-ProX-Lite		
	Acc	Tok	Acc/kT	Acc	Tok	Acc/kT
English	0.944	265	3.57	0.379	565	0.67
Chinese	0.848	266	3.18	0.418	570	0.73
Spanish	0.884	287	3.09	0.284	592	0.48
Russian	0.872	291	2.99	0.309	591	0.52
French	0.839	286	2.93	0.302	586	0.51
German	0.843	294	2.87	0.251	596	0.42
Japanese	0.786	284	2.77	0.319	582	0.55
Thai	0.788	345	2.28	0.188	613	0.31
Swahili	0.447	296	1.51	0.163	601	0.27
Bengali	0.601	401	1.50	0.113	624	0.18
Telugu	0.318	469	0.68	0.094	604	0.16

Table 1: Per-language single-trace accuracy, mean generated tokens per trace, and token efficiency (accuracy per thousand generated tokens), over the full pools. Accuracy and cost move in opposite directions.

Corpus	Traces	Samples, one language		One trace per language		Samples — languages	
		Ceiling	Vote	Ceiling	Vote	Δ ceiling	Δ vote
MGSM	2	0.972	0.941	0.965	0.941	0.007 [0.000, 0.015]	0.000 [−0.003, 0.003]
	4	0.988	0.962	0.984	0.949	0.004 [−0.001, 0.010]	0.013 [−0.002, 0.028]
	8	0.992	0.966	0.989	0.956	0.003 [0.001, 0.007]	0.010 [−0.009, 0.029]
MMLU-ProX	2	0.487	0.426	0.464	0.393	0.023 [−0.077, 0.124]	0.032 [−0.064, 0.129]
	4	0.591	0.462	0.617	0.447	−0.026 [−0.119, 0.069]	0.015 [−0.080, 0.109]
	8	0.685	0.487	0.704	0.453	−0.018 [−0.119, 0.085]	0.034 [−0.066, 0.137]

Table 2: Two sources of diversity at a *matched trace budget*: resampling one language at temperature, and one trace from each of the k strongest languages. Ceiling is the per-item trace-level oracle over the k traces; Vote is majority over the same traces. Differences are paired item bootstraps, 4,000 replicates, positive where samples lead. On MGSM the language ballot buys no ceiling and no vote that resampling does not, at $1.09\times$ the tokens at $k=8$; on MMLU-ProX-Lite every interval covers zero at $n=124$, so the two are not separable at a matched budget there either. Figure 1 plots the same paired quantities.

5.4 Aggregate error independence does not explain the failure

Were the languages to fail on the same items, a vote would have no independent evidence to aggregate and the preceding result would follow trivially. They do not. Mean pairwise agreement on item-level correctness is 0.718, and conditioned on one of two languages being wrong, both are wrong on only 0.215 of items on MGSM. Failures are largely disjoint, which is why the oracle is so high.

Appendix C shows this pair by pair. On MGSM the structure is uniform: no pair fails together often, including pairs of two weak languages. The independence is therefore a property of the ensemble and not an artifact of pairing everything against a strong English column.

Inequality among the voters is a second contributing factor. Table 1 shows single-trace accuracy from 0.944 in English down to roughly a third of that in Telugu, with token efficiency spanning $5.3\times$ on MGSM and $4.7\times$ on MMLU-ProX-Lite. Reweighting the ballot by measured competence

Rule (one trace per language)	MGSM	MMLU-ProX
Best single language	0.935	0.383
Uniform majority vote	0.946	0.340
Independence model prediction	1.000	0.820
Accuracy-weighted vote	0.945	0.337
Log-odds-weighted vote	0.953	0.341

Table 3: Aggregation rules over one trace per language. The independence model draws each language correct with its measured marginal accuracy and otherwise assigns it a distinct wrong answer; its over-prediction of the measured uniform vote is what identifies the missing structure as item-level rather than pairwise. Weights are fit on the validation half.

does not repair the vote, however (§5.5), which is what directs the diagnosis toward item-level structure rather than toward voter weights.

5.5 An independence model over-predicts the vote

Table 3 compares rules that consume exactly one trace per language. The independence model draws each language correct with its measured marginal

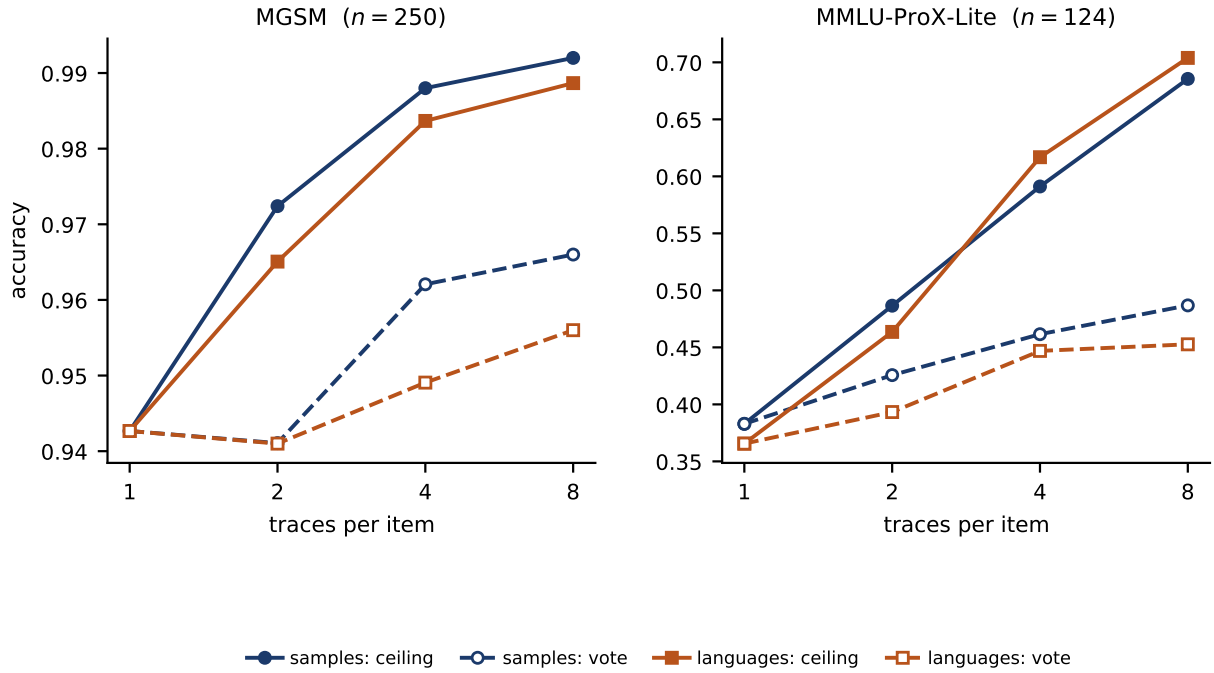


Figure 1: Ceiling (solid) and majority vote (dashed) against the number of traces per item, for two sources of diversity: temperature resampling within one language (blue) and one trace from each of the k strongest languages (orange). On MGSM the two are on top of each other and the language ballot is the more expensive; on MMLU-ProX-Lite the language ceiling pulls ahead while its vote falls further behind, which is the gap the rest of the paper explains.

accuracy and otherwise assigns it a distinct wrong answer. It predicts 1.000 on MGSM where the measured uniform vote returns 0.946, and 0.820 against 0.340 on MMLU-ProX-Lite — an over-prediction of 5.4 and 48.0 points respectively. The marginals and the pairwise statistics are therefore jointly insufficient to predict aggregation behaviour, and the missing structure is at the item level.

Weighting the ballot by validation accuracy returns 0.945 and by log-odds 0.953 on MGSM, against 0.946 uniform and 0.935 for the best single language: reweighting moves the result by less than a point and does not overtake the strongest language. A rule that assigns fixed per-language weights cannot recover the headroom, because the languages that are wrong on a given hard item are not a fixed subset.

On MMLU-ProX-Lite the agreement structure differs sharply: both-wrong-given-either-wrong is 0.611, roughly three times the MGSM value, with mean agreement 0.685. On harder items the languages do fail together more often, and the oracle correspondingly sits closer to what a rule could reach.

5.6 Per-item difficulty is heterogeneous

Figure 2 isolates the structure the independence model omits. On MGSM, 0.172 of items are answered correctly by every language and 0.016 by none, leaving 0.812 in between; on MMLU-ProX-Lite the shares are 0.088, 0.328 and 0.584. Items are not exchangeable draws with a common per-language success probability, which is what the model assumes and what its marginals encode.

The consequence for voting is direct. On items every language solves, a vote is correct but redundant, since a single language would have sufficed. On items almost none solves, a vote is wrong regardless of ballot size. Aggregation can only change the outcome in the intermediate region, and there the wrong answers concentrate: among items where at least three languages err, the single most common wrong answer is held by 0.464 of the erring languages on MGSM and 0.549 on MMLU-ProX-Lite. A plurality of wrong languages agreeing on one wrong answer is exactly the configuration a majority vote cannot survive, and it is invisible in both the marginal accuracies and the pairwise statistics.

Per-language cost also splits into a language component and a failure component: incorrect

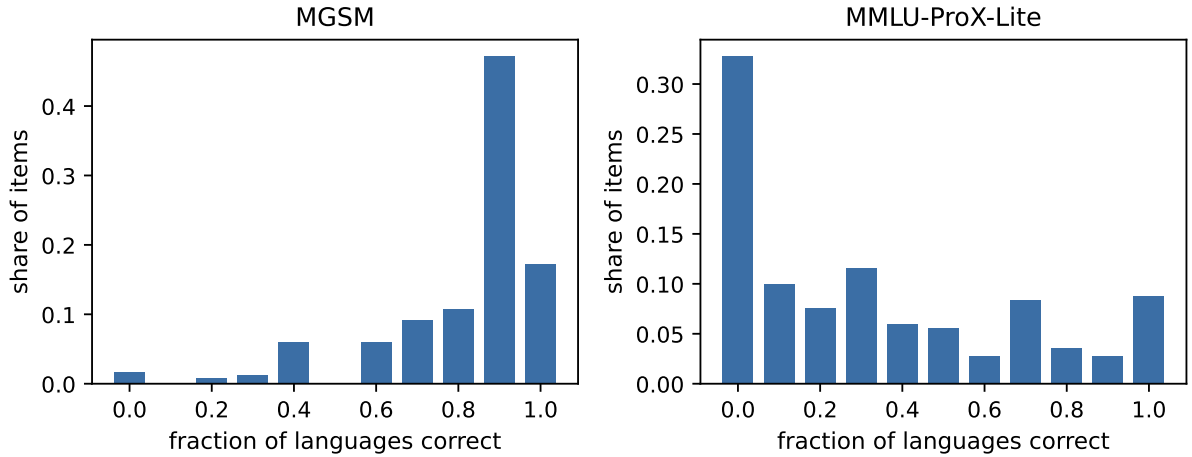


Figure 2: Distribution over items of the fraction of languages answering correctly. Mass at the right edge is items every language solves; mass at the left is items almost none solves. An independence model with per-language marginals has no parameter for this axis.

traces run $1.23\times$ – $1.58\times$ longer than correct ones in every language, so roughly 22% of Telugu’s apparent $1.77\times$ token penalty over English is composition rather than the language being costlier to reason in (Appendix D).

5.7 The size of the induced routing problem

Because the obstacle is dispersion rather than dependence, the constructive response is to select rather than to broaden. Restricting a perfect router to the m strongest languages (Table 5, Appendix A), two languages already reach 0.980 on MGSM — 80% of the headroom over the best single language — and three reach the full 11-language ceiling to three decimals, so the routing problem there is a binary or ternary decision per item. On MMLU-ProX-Lite the oracle instead climbs gradually from 0.578 to 0.672 and a router would need a wider set; the two corpora bracket the design space rather than agreeing.

5.8 A floor beneath the ceiling

An oracle is an upper bound and says nothing on its own about what is reachable. Table 4 places three selectors beneath it, each computed from the released pool with no access to the gold answer: the best single language’s self-consistent answer; that answer unless the second-ranked language disagrees and is more self-consistent; or the answer of whichever language is most self-consistent on the item. Rankings are fit on the validation half.

None beats the best single language. Self-consistent English returns 0.968; top-2 agreement 0.968 at twice the tokens, the most-self-consistent

rule 0.936, and the uniform eleven-language vote 0.960, against an oracle of 0.984. The routing headroom is 1.6 points and no signal we tried recovers any of it; a perfect router over two languages would reach 0.976 on this half, half of it. §5.7’s 0.980 is the same oracle against a single-trace baseline over the full pool; here the baseline is self-consistent, so less headroom sits above it.

MMLU-ProX-Lite is where routing is worth pursuing: best single 0.416 against an oracle of 0.528, a gap of 11.2 points, seven times the MGSM gap, with the uniform vote at 0.424. Self-consistency is a weak signal exactly where the model is weak, which is where the headroom is. We report this as a floor rather than a method: it is what a selector should be scored against, and it says the problem is open rather than easy.

Which languages carry answers no other language supplies is reported in Appendix E. The ballot sweep repeated on MMLU-ProX-Lite is reported in Appendix F; its shape is the same and per-item language coverage in that pool is partial, so we draw no comparative claim from it.

5.9 Registered predictions

We registered six predictions before running these re-scores. Two missed, and we can now say why each missed rather than only that it did.

The first miss is the result itself: we predicted cross-lingual aggregation would beat English-only by +1 to +3 points under sample matching and measured -0.7 and -5.1 . That prediction was an inference from the aggregate error statistics, which

Rule	Acc	Tok/item	vs. best single
Best single language (self-consistent, $k=8$)	0.968	2,167	—
Router: top-2 agreement	0.968	4,538	+0.000 [+0.000, +0.000]
Router: most self-consistent language	0.936	28,415	−0.032 [−0.072, +0.008]
Uniform vote, all 11 languages	0.960	28,415	−0.008 [−0.040, +0.016]
Oracle over 2 languages	0.976	4,538	+0.008 [+0.000, +0.024]
Oracle over 11 languages (<i>ceiling</i>)	0.984	28,415	+0.016 [+0.000, +0.040]

Table 4: MGSM, held-out problem ids. A floor beneath the ceiling: three selectors that never see the gold answer, against the best single language and the oracle. No selector beats the best single language, so the 1.6 points between it and the ceiling are open. Intervals are paired item bootstraps, 2,000 replicates.

are favourable, and those statistics average over items whose difficulty is not exchangeable: the independence they describe sits on items every language already answers correctly and vanishes on the items that decide the score. Predicting a gain from marginal and pairwise statistics is exactly what the independence model in Table 3 does, and it over-predicts by 5.4 and 48.0 points. The prediction failed for the reason the model fails.

The second is an instrument error on our side rather than a fact about the world. We predicted a $1.25\text{--}1.45\times$ token subsidy for uniform eleven-language sampling and measured $1.10\times$ and $1.04\times$ — under the fixed- k protocol, which never draws the three most expensive languages the prediction was about. The registered and measured quantities are not the same estimator, so we report this as an instrument difference rather than a failed prediction; under coverage the eleven-language ballot costs $13.1\times$ a single language. We predicted pairwise error agreement above 0.60, which held on MMLU-ProX-Lite (0.611) and failed on MGSM (0.215); that failure is what identifies competence dispersion rather than error correlation as the obstacle, and is the reason the paper’s conclusion is a routing claim rather than an ensembling one. The oracle prediction and the length-by-correctness prediction both held on both corpora.

6 Discussion

Two things follow for how inference-time methods are compared. The first is accounting: a trace is not a unit of compute, so comparisons should be reported at matched tokens. The subsidy is small here ($1.10\times$) but it is not zero and it grows with the fraction of the ballot spent on expensive scripts.

The second is larger and is not specific to languages. Every method in this family buys diversity, and the cheapest diversity available is to ask the same model the same question again. A new source

of diversity therefore has to clear that bar, and clearing it is a claim about *coverage* — whether the expensive source reaches items the cheap one never reaches — which is a statement about the ceiling and not about the vote. On MGSM cross-lingual generation does not clear it at any budget we can measure. On MMLU-ProX-Lite it clears it by a wide margin and the vote still loses, because the coverage arrives on exactly the items where the languages that hold it are outnumbered. Reporting the ceiling of the cheap policy beside the expensive one costs nothing when the traces already exist, and it is the check that separates the two cases.

Section 5.8 builds three such selectors from the obvious per-item signals — self-consistency within a language, and agreement between a language pair — and none of them beats the best single language. What a router needs on the items that carry the headroom is therefore evidence these pools do not contain, and finding it is the natural subsequent use of this artifact.

7 Conclusion

Compared at a matched number of samples, cross-lingual aggregation looks like a way to reach more of what a multilingual model knows. Compared at a matched budget, on the corpus where the model is strong it reaches nothing that resampling one language does not: the eighty traces from ten further languages solve 0 MGSM items that English’s own eight do not, at $1.09\times$ the cost. Where the model is weak the coverage is real — 26 MMLU-ProX-Lite items are solved only outside English — and voting still returns less than monolingual voting at the same budget, because the languages holding that coverage are outnumbered on the items that carry it. The obstacle is not that the languages agree in their errors but that they differ in competence, and no selector we could compute from the released traces beats the best single language.

8 Limitations

Our monolingual arm varies traces by temperature under one fixed prompt, so it isolates sampling diversity but not prompt diversity: an English ballot built from several paraphrases of the item would be a third source of diversity, cheaper than translation and not measured here, and it is the control we would run next. Results are for one open-weight model at the 8B scale, so the magnitude of the competence dispersion we measure is a joint property of this model’s tokenizer and training mixture and will differ for others. Both corpora are item-parallel translations rather than natively authored items, so per-language accuracies describe translated evaluation material and may understate the competence of languages whose native content is unlike translated English. We aggregate over final answers only, so a rule that inspects intermediate reasoning could exceed the answer-level ceiling reported here. Comparative policy claims are made on both corpora, but the MMLU-ProX-Lite arm rests on $n=124$ items, wide enough that every matched-budget interval there covers zero; what that corpus supports is the coverage statement over the full pool, not a separation at matched budget. And MGSM is close to saturated for this model at 0.944 in English, so the headroom available to any policy on it is correspondingly small.

References

- Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2026. Cross-lingual self-consistency for multilingual reasoning with language models. *arXiv preprint arXiv:2606.01464*.
- Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Sara Rajaei, Rochelle Choenni, Ekaterina Shutova, and Christof Monz. 2025. Best-of-l: Cross-lingual reward modeling for mathematical reasoning. *arXiv preprint arXiv:2509.15811*.
- Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. 2024. [Empowering multi-step reasoning across languages via program-aided language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12171–12187, Miami, Florida, USA. Association for Computational Linguistics.
- 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. 2023. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, and Yun Xing. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.
- Zhiwei Yu, Tuo Li, Changhong Wang, Hui Chen, and Lang Zhou. 2025. Cross-lingual consistency: A novel inference framework for advancing reasoning in large language models. *arXiv preprint arXiv:2504.01857*.

A Oracle by number of languages available

Languages available	MGSM	MMLU-ProX
top-1	0.964	0.520
top-2	0.980	0.578
top-3	0.984	0.578
top-4	0.984	0.610
top-5	0.984	0.628
top-6	0.984	0.647
top-7	0.984	0.616
top-8	0.984	0.648
top-9	0.984	0.656
top-10	0.984	0.656
top-11	0.984	0.672

Table 5: Oracle accuracy when a perfect router may choose among the m strongest languages. Most of the ceiling is reachable from a small set, so the routing problem this poses is small.

B The ballot sweep on MGSM

Langs. avail.	Fixed $k=8$		Coverage	
	Acc	Tok	Acc	Tok
top-1	0.950	2,173	0.929	270
top-2	0.956	2,267	0.934	568
top-3	0.960	2,304	0.947	867
top-4	0.945	2,279	0.942	1,139
top-5	0.950	2,307	0.944	1,439
top-6	0.948	2,308	0.949	1,736
top-7	0.958	2,360	0.945	2,089
top-8	0.949	2,376	0.949	2,374
top-9 [†]	0.951	2,374	0.950	2,777
top-10 [†]	0.949	2,373	0.943	3,082
top-11 [†]	0.951	2,373	0.944	3,552

Table 6: MGSM ballot sweep. *Fixed k* is the prior protocol at $k=8$; *Coverage* draws one trace per language. Under the prior protocol a fixed k round-robin cannot reach more than k languages, so rows marked [†] vote over only the eight highest-ranked languages however many are nominally available — visible in the flat token column. Sampling one trace per language instead lets every language enter, and the token cost then rises with the ballot as it should. Accuracy does not rise under either protocol.

Adding languages to the ballot does not raise accuracy under either protocol.

C Pairwise error agreement

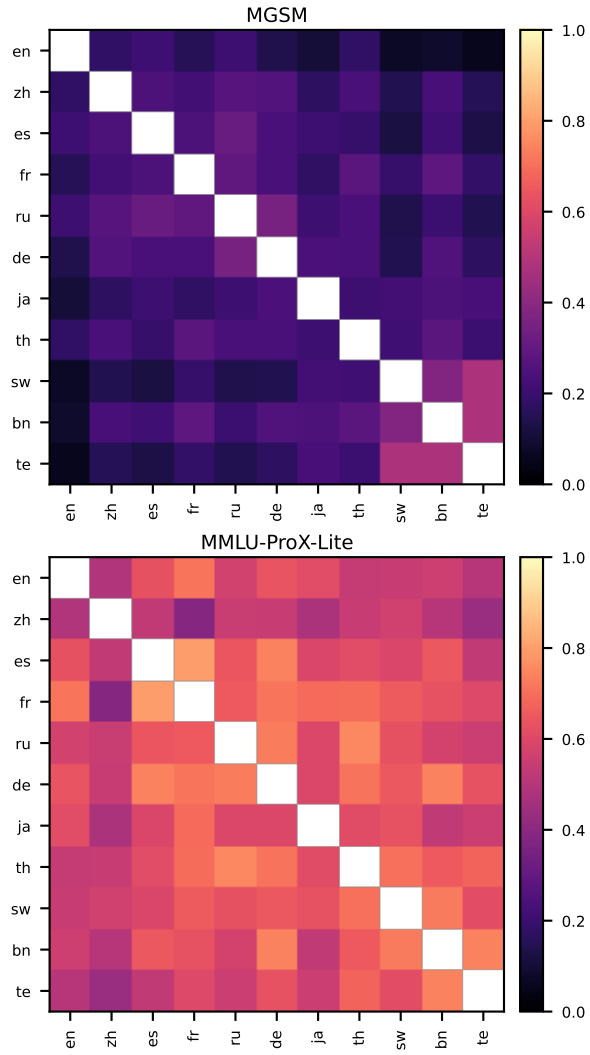


Figure 3: Pairwise $\Pr[\text{both wrong} \mid \text{either wrong}]$ for all language pairs, on a shared 0–1 scale. Dark cells are pairs whose failures are disjoint. MGSM (top) is uniformly dark — its values span roughly 0–0.4 — while MMLU-ProX-Lite (bottom) is lighter, i.e. on harder items the languages increasingly fail together.

D Trace length by correctness

Language	Tok (correct)	Tok (wrong)	Ratio
en	256	405	1.58
zh	254	333	1.31
es	272	402	1.48
fr	270	371	1.37
ru	275	404	1.47
de	279	372	1.33
ja	264	355	1.34
th	316	451	1.42
sw	262	323	1.23
bn	347	482	1.39
te	400	501	1.25

Table 7: MGSM. Incorrect traces are longer than correct ones in every language. Part of the per-language cost in Table 1 is therefore composition: weaker languages fail more often, and failures run longer.

The per-language cost in Table 1 averages over correct and incorrect traces, and those differ: conditioning on correctness (Table 7, Appendix D), incorrect traces are longer in every language on MGSM by $1.23\times$ to $1.58\times$. That separates two things the raw cost conflates. Telugu costs $1.77\times$ English’s tokens per trace overall but only $1.56\times$ when both are correct, so roughly 22% of the apparent penalty is composition — Telugu is wrong far more often, and wrong traces run long — rather than the language being intrinsically more expensive to reason in. Cost accounting that ignores this attributes to the script what belongs to the failure rate.

E Which languages carry unique answers

Language	MGSM	MMLU-ProX
en	0	8
es	0	4
fr	0	2
de	0	7
ru	0	7
zh	0	11
ja	0	4
th	0	7
sw	0	2
bn	0	0
te	0	4
Items solved by exactly one	0	56

Table 8: Items on which a language is the *only* one to answer correctly. A language with a non-zero count carries headroom no other language supplies, and is therefore worth a router’s attention independently of its marginal accuracy.

The routing measurement in §5.7 says a small language set suffices on MGSM and a larger one is needed on MMLU-ProX-Lite. Table 8 supplies the reason. On MGSM, 0 items are solved by exactly one language: every item any language solves is solved by at least two, so narrowing the candidate set discards no reachable item and the two-language ceiling coincides with the eleven-language one. On MMLU-ProX-Lite, 56 of the solved items have exactly one solver, and those solvers are spread over 10 different languages including the weakest ones in Table 1.

This qualifies the routing result rather than extending it. Breadth is redundant on items a model finds easy and load-bearing on items it finds hard, so the number of languages worth consulting is a function of item difficulty rather than a constant of the model. It also identifies a cost of the narrow-set strategy that a marginal-accuracy ranking cannot see: a language may contribute nothing on average and still be the only source of a correct answer on a specific hard item.

F The ballot sweep on MMLU-ProX-Lite

Languages in the vote	$k=8$	$k=16$
top-1 (best single)	0.408	0.431
top-2	0.458	0.468
top-3	0.435	0.435
top-4	0.427	0.430
top-5	0.422	0.434
top-6	0.443	0.460
top-7	0.390	0.400
top-8	0.395	0.411
top-9	0.396	0.404
top-10	0.378	0.396
top-11 (uniform)	0.383	0.397

Table 9: MMLU-ProX-Lite. Same sweep as Table 6. Reported for completeness; per-item language coverage in this pool is partial (§4), so these cells are not paired across policies and we draw no comparative claim from them.

Table 9 repeats the ballot sweep on MMLU-ProX-Lite. The pattern is the same in shape — accuracy does not increase with ballot size, and the uniform eleven-language vote (0.383) is not the best cell — but per-item language coverage in this pool is partial, so these cells are not paired across policies. A paired bootstrap on the 31 items where the comparison is well defined gives +4.26 points for a two-language ballot over the uniform one, with a 95% interval of $[-1.94, +11.23]$. We therefore report the sweep for completeness and make no comparative claim from this corpus. The per-language statistics, the oracle and the agreement structure do not require pairing and are reported without this caveat throughout.