BharatMath: Culturally Regenerated Mathematical Reasoning for Twelve Indic Languages

Anonymous authors
Paper under double-blind review

Abstract

Mathematics benchmarks for Indian languages are translations: they carry the dollars, miles and bake sales of their English sources into every language, they are small, and their English halves are saturated. We introduce **BharatMath**, 88,066 word problems in twelve Indic languages that are *regenerated* rather than translated. An executable formula is extracted from each English source problem, fresh operands are sampled and the gold answer is computed in code, so no language model can alter the stored contract; whether the generated question expresses it is audited separately. The question is then rebuilt around a record from a **Context Bank**: a finite, typed, enumerable relation of 782 curated cells that allocates regional units and scenarios by language and mathematical domain, so that what the corpus is *about* is specified before generation and auditable after it. On identical contracts rated blind, the whole pipeline clears a cultural-plausibility criterion on 52% of items against 0.7% for literal translation; the bank's own contribution is $+9$ points of regional appropriateness, no significant change in cultural plausibility, and -5 points of mathematical equivalence. Golds are machine-checkable (99.94% reproduce from the stored formula); a two-family gold-blind LLM audit labels 8,778 of the 8,860 public test items (82.7% fully valid, 13.2% ambiguous, 2.2% with an altered operand, 1.9% with a differing answer), and on the valid core the model ranking is identical and the intervention effects below persist. Seven open-weight models span five points on MGSM-English but 11 to 60 points between their best and worst Indic language at matched decoding, and regenerated Bengali orders them differently from translated Bengali. Because every item is a contract plus a rendering, the benchmark also supports single-factor interventions, which locate six gaps: question comprehension (up to 52 points), numeral script (up to 34), a natural-language execution component for one model family, sampling-accessible headroom, the cost of forcing native-script thought under a fixed token budget (4–20 points, mostly token-cap hits for two of three models), and distractor robustness. Re-minting operands costs 8.5 points on items that keep their source's numbers and 0.7 (n.s.) on items whose numbers were already resampled. We release the corpus, the bank, the lineage of every item, and keep a private held-out split.

1 Introduction

Mathematical word problems are the most widely used probe of multi-step reasoning in language models, and for Indian languages the available probes are *translations*. MGSM (Shi et al., 2022) renders 250 GSM8K problems (Cobbe et al., 2021) into ten languages, two of them Indic; MSVAMP (Chen et al., 2023a) does the same for SVAMP (Patel et al., 2021). Translation preserves the gold answer and everything else: the bake sale, the dollar, the mile and the 15% tip survive into Bengali and Telugu. The translated item tests reading of a foreign scene in a native script, and cultural framing is known to move accuracy with the mathematics held fixed (Tomar et al., 2025); translationese is a learnable shortcut (Koppel & Ordan, 2011; Artetxe et al., 2020); and the source items are public and heavily trained on (Mirzadeh et al., 2025; Zhang et al., 2024).

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

Table 1: Indic mathematics and culture benchmarks. MILU is multiple-choice knowledge, not free-response arithmetic; A-Loc. is the localisation framework of Azime et al. (2026). Evaluation counts are MGSM's 250×10 (500 items in its two Indic languages), the 1,319 GSM8K items DIWALI adapts, and BharatMath's 8,860 public plus 8,870 private test items. "Shallow" denotes surface substitution; "answer recomputed" denotes evaluating the extracted formula on newly sampled values.

	MGSM (2022)	**MILU** (2025)	**DIWALI** (2025)	**A-Loc.** (2026)	**BharatMath**
Indic languages (of 12)	2	11	—	—	**12**
Scenario-level grounding	×	partial	shallow	shallow	✓
Sub-regional control	×	×	× (skewed)	×	✓
Answer recomputed	×	n/a	×	×	✓
Item lineage released	×	×	×	×	✓
Evaluation items	2,500	85k	1,319	∼1k	**17,730**

The obvious alternative, asking a strong model to "localise" each problem, fails differently. Post-training pushes generators toward typical continuations (Zhang et al., 2025; Wu et al., 2025a; Wright et al., 2025), and prompted cultural adaptation makes surface substitutions and favours high-resource states (Sahoo et al., 2025). *What* the corpus is about becomes an accident of the generator's prior, invisible until the corpus is sampled and impossible to audit afterwards.

BharatMath addresses both failures with one design decision: separate what must never change from what should. The mathematics of each item — an executable formula, freshly sampled operands and a gold answer computed in code — is frozen in a *contract* that no language model can edit; whether the generated question still expresses that contract is a separate, audited property (§4). The cultural surface — units and scenario — is allocated by a **Context Bank**, a finite typed relation keyed by language and mathematical domain, whose key distribution is fixed before generation and verified by enumeration afterwards. Table 1 positions the result. We contribute:

- **A construction.** Context Banks make cultural allocation explicit, reusable and inspectable (§3). Rated blind on 600 identical contracts, the whole pipeline is culturally plausible on 51.7% of items against 0.7% for literal translation; the bank itself adds $+9.1$ points of regional appropriateness at -5.3 points of mathematical equivalence (§4).
- **A benchmark.** 88,066 items, twelve languages, fifteen domains, lineage-disjoint train, public-test and private-test splits, with gold integrity, question validity and ranking stability measured rather than assumed (§4). At matched decoding, language spreads run from 11 to 60 points, and the Bengali ranking differs from MGSM-Bengali's (§5).
- **A diagnosis.** Paired single-factor interventions across seven models locate six evaluation gaps, each with an effect size and interval (§6), and an operand re-mint with a built-in control stratum measures a loss of 8.5 points on items that keep their source's operands against 0.7 (n.s.) on the rest (§7).

2 Related work

Multilingual and Indic evaluation. MGSM and MSVAMP translate English problem sets; MILU (Verma et al., 2025), IndicGenBench (Singh et al., 2024), INCLUDE (Romanou et al., 2025) and Global-MMLU (Singh et al., 2025) cover knowledge and generation rather than multi-step arithmetic. Translated sets correlate poorly with human judgements in the target locale (Wu et al., 2025b). Re-instantiating a multilingual set is not new: MGSM-Pro (Xu et al., 2026) gives five instantiations of each MGSM question in nine languages by varying names, digits and irrelevant context; PolyMath (Wang et al., 2025) and MMATH (Luo et al., 2025) scale multilingual mathematics to eighteen and ten languages, at most two of them Indic. What BharatMath adds is coverage of twelve Indic languages, explicit allocation of scenarios, per-item lineage, and a paired intervention design. **Cultural adaptation.** DIWALI (Sahoo et al., 2025), DOSA (Seth et al., 2024) and localisation frameworks (Azime et al., 2026) supply cultural items or prompt a model to substitute them; Tomar et al. (2025) show that re-templating GSM8K into other regions costs up to 19 points. None controls the *distribution* of contexts or recomputes the answer. **Contamination-resistant mathematics.** GSM-Symbolic (Mirzadeh et al., 2025) and GSM1k (Zhang et al., 2024) report that models lose accuracy

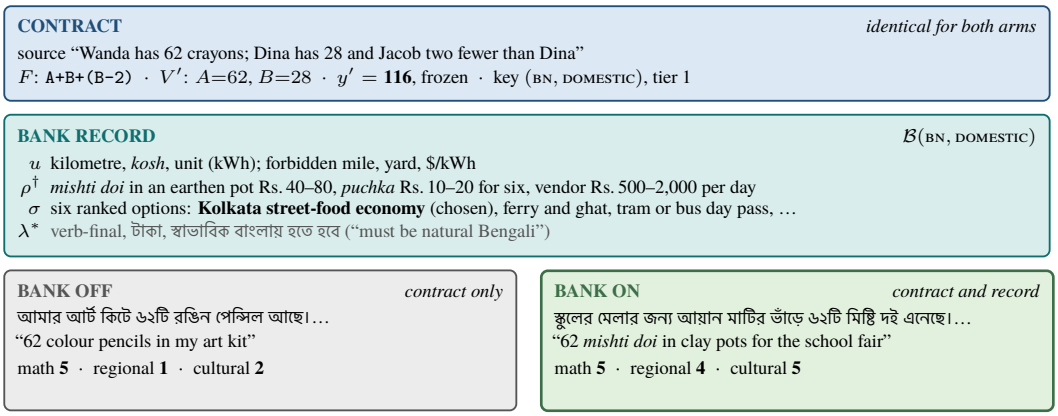


Figure 1: One contract, two generations. Both arms receive the same formula, sampled values and frozen answer; only the second also receives the curated record for its (L, D) cell, of which the evaluated build supplies u and σ (†advisory, *retrieved but unused). The generator took the record's street-food option, down to the earthen pot. A blind rater scored each question 1–5. This pair illustrates the mechanism and is not a typical effect size: over all 600 contracts the bank's measured gain is on the regional axis (Figure 3).

Table 2: What the bank knows. Five of the 438 unit cells and 223 scenario rows, all for land, trade and household domains. × marks units the bank *forbids* for that language: the same word can be required in one language and wrong in another. All twelve languages are in Table 5.

Language	u: units preferred, and forbidden	σ, ρ: scenario and magnitudes
Hindi	*bigha*, *biswa*, *seer* (933 g), *maund*; × acre, hectare	UP cane price Rs. 370/quintal; brick-kiln piece rate Rs. 450–650 per 1,000 bricks
Punjabi	*kanal*, *marla* (30.25 sq yd), *ghumao* = 8 kanal = 1 acre; × bigha (UP variant), katha	*arhatiya* commission 2.5% + Rs. 2.50/quintal; stubble fine Rs. 2,500/acre
Bengali	*bigha* (14,400 sq ft = 20 *katha*), *bhori* (11.66 g gold); × guntha, kanal, sovereign	jute MSP Rs. 5,335/quintal at 10–16 *maund* per *bigha*
Tamil	*cent*, *ground* (2,400 sq ft), *kuli* (72 sq ft), *padi* (≈1.25 kg); × bigha, guntha	Amma canteen *idli* Rs. 1, sambar rice Rs. 5; Kanchipuram weaver Rs. 300–600/day
Malayalam	*cent* (1/100 acre), *pavan* (8 g gold), *para* (8–13 kg); × tola, bigha, guntha	RSS-4 rubber Rs. 170–210/kg; KSFE *chitty*, foreman commission 5%

on re-instantiated GSM8K-style items, though a mixed-model re-analysis finds the drop significant for only 8 of 20 models and partly explained by larger resampled integers (Długosz et al., 2026); we therefore control magnitude and cluster by item in §7. The public/private test design follows Truong et al. (2026). **Diagnostics.** Our instruments are program-aided solving (Gao et al., 2023; Chen et al., 2023b), pass@k (Wang et al., 2023; Chen et al., 2021), budget forcing (Muennighoff et al., 2025), distractors (Shi et al., 2023), cross-lingual thought (Huang et al., 2023; Etxaniz et al., 2024) and tokeniser analysis (Petrov et al., 2023; Ahia et al., 2023).

3 Context Banks and the construction

Definition. A **Context Bank** is a finite typed relation $\mathcal{B} : (L, D) \mapsto (u, \rho, \sigma, \lambda)$, where L is the target language and D the mathematical domain, optionally refined by a sub-region. The record specifies admissible and forbidden units u, plausible magnitudes and rates ρ, a scenario specification σ and language-structural guidance λ. These fields are the problem's mutable cultural context; formula, sampled values and gold are stored separately and cannot be changed by a generator; units are free text without dimensional typing, so unit fidelity is audited (§4), not guaranteed. Regional refinement is part of the key: HI/Land–Western UP and HI/Land–Purvanchal return different *bigha* conversions, and *bigha* is a preferred land unit in cells of seven languages, names three different areas, and is forbidden in the school-geometry cells of all twelve (Table 2; counts are over (L, D) cells, so they overlap). A generic "use an Indian setting" does not say which convention to follow; the bank records the choice once.

Why a finite relation. Because $\mathcal{B}$ is finite, its key distribution can be specified before generation and checked afterwards against an enumerable key set. The conditioning supplied to the generator

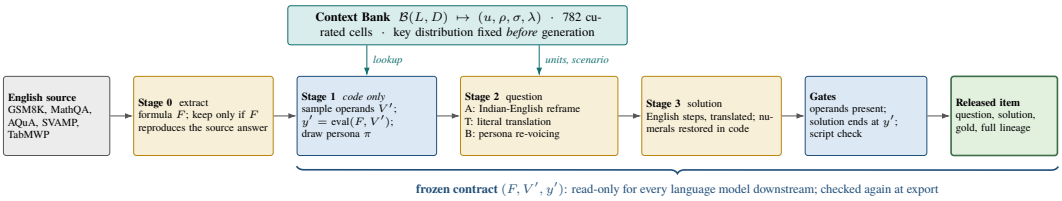


Figure 2: The construction. Language models (amber) write text; code (blue) owns the mathematics. The Context Bank (teal) decides what each item is *about*, by language and domain, before any text is generated. Appendix B gives the full specification and a worked example.

becomes checkable by enumeration rather than by inspecting sampled outputs; a repeated key returns the same curated record; and a wrong fact is corrected in one cell, with the items generated from it listed from the lineage for re-checking (Appendix C.4). The released bank has **782 cells** in three layers: units (438), world scenarios and magnitudes (259) and language-structural guidance (85), authored offline and reconciled against public reference data such as crop prices, state wages, tax slabs and transport fares. 250 of 438 unit cells and 40 of 223 scenario rows name a sub-region. The method evaluated in this paper is narrower than the definition: u and σ condition generation, ρ reaches the generator only as advisory guidance, so released operands come from blueprint ranges and not from bank magnitudes, and λ is retrieved but never enters a prompt (Appendix B). Figure 1 shows one contract generated with and without its record.

Pipeline. Four stages consume the bank while keeping the mathematics out of reach of the language models (Figure 2). *Stage 0* extracts an executable formula F from an English source problem drawn from GSM8K, MathQA (Amini et al., 2019), AQuA (Ling et al., 2017), SVAMP and TabMWP (Lu et al., 2023), and keeps it only if it reproduces the source answer. *Stage 1* is code: it samples fresh values V' (a source value survives only where Stage 0 emitted a degenerate range; §7), computes $y' = \text{eval}(F, V')$, draws a persona from Nemotron-Personas-India (NVIDIA, 2025) and retrieves $\mathcal{B}(L, D)$; a job with no curated record is discarded, never run without context. *Stage 2* realises the question in three calls: an Indian-English reframe around the bank record, a literal translation, and a persona re-voicing. *Stage 3* writes a step solution and restores numerals in code. Export requires the generated solution to end at y' and the question to pass the numeral gate of Table 6; the model readers of §4 are audits, not export gates, and the final text is not digit-checked, so operands survive as digits in 97.2–98.8% of questions, not all. Every released item carries its lineage: source id, formula, operands, bank key, persona and every intermediate text.

The corpus. The 88,066 items cover fifteen domains (commercial arithmetic 31.1%, algebra 21.0%, motion 14.8%); 64.6% need more than one step and the median is three operations. Items name 80 to 444 distinct districts and 986 to 1,403 distinct occupations per language (Figure 9). The 70,336 / 8,860 / 8,870 train / public-test / private-test split is grouped by source lineage: no lineage or numeric-fingerprint group crosses a boundary, and rows linked to public MathQA, TabMWP and SVAMP test items are quarantined in train.

4 Does the construction work?

The bank changes where items are set. A pre-registered, blind-rated study compares four generation methods on 600 released-item contracts (50 per language), with formula, sampled values and gold fixed across methods (Figure 3, left). Literal translation clears the regional criterion on *zero* of 600 items and direct generation in the Indic language on 22.8%; the two pipeline arms are 30 to 64 points above both on the cultural axes. Isolating the bank from the pipeline, C4−C2 yields $+0.091$ regional appropriateness ($p \leq 0.002$, Holm-corrected, gate-matched; $+0.113$ [$+0.077$, $+0.150$] ungated and $+0.114$ on first attempts only, so gate history does not produce it), $+0.013$ cultural plausibility (n.s.) and -0.053 mathematical equivalence ($p = 0.010$): the bank's measured effect is regional appropriateness, bought at a fidelity cost, and not cultural plausibility (Appendix C). A twelve-arm replication on 600 *fresh* contracts with a different judge (Figure 3, right) shows which part of the bank matters. Conditioning costs no first-attempt validity with this generator and 7–13 points with Qwen3-235B. The bank lifts the share of items a blind judge places in the right state from 0.06 to

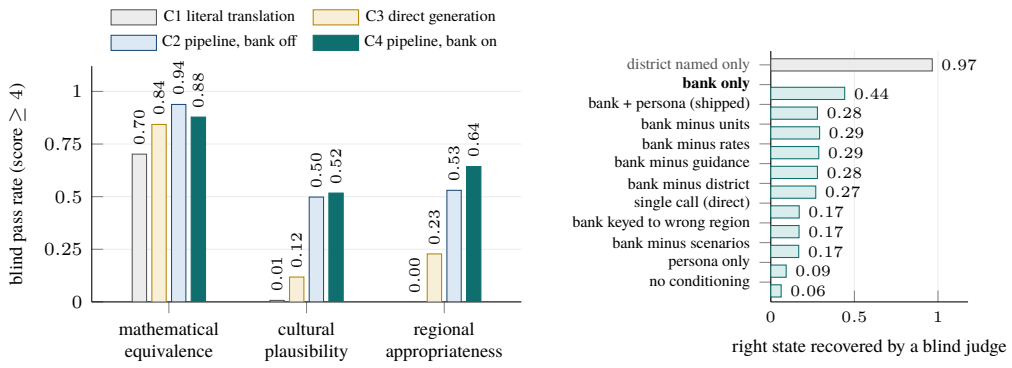


Figure 3: What the pipeline and the bank each contribute. *Left:* four generation methods on 600 identical contracts, rated blind on a 1–5 rubric; only the question text differs. Translation keeps most of the mathematics and none of the setting; the pipeline keeps both, and the bank adds regional appropriateness (+9.1 points gate-matched, $p \leq 0.002$) at a measured cost in mathematical equivalence. *Right:* a twelve-arm replication on 600 fresh contracts (21,600 first-attempt generations; blind Qwen3-235B judge). Removing the scenario layer or keying the bank to another region undoes the gain; the other layers carry none (state recovery measures geographic cues, not cultural fit). The grey bar simply prints the district name and is a ceiling reference, not a method.

0.28–0.44 (recoverable geographic cues, not cultural appropriateness); removing the scenario layer or keying the bank to another region costs 11 points, and no other layer carries a measurable share. A second generator, Qwen3-235B, replicates every grounding contrast (Table 9, Appendix C.3–C.4).

The gold is machine-checkable. Re-evaluating the stored formula on the stored operands reproduces the released gold on 99.94% of items; the final boxed value of the released solution equals the gold in 99.98% of traces. An independent re-derivation of the formula from the source problem is functionally identical to the stored one on 88.0% of 5,764 sources and the audit flags 2.7% of them (Appendix D).

Does the question express the contract? This is the binding constraint of any regenerated benchmark, so we measure it two ways, both gold-blind. (i) A two-family audit with blind adjudication runs over the *whole* public test split and labels 8,778 of its 8,860 items (the other 82 lack a verdict from one family or stay undecided, and ship unlabelled): 82.7% of labelled items are fully valid; the dominant defect is ambiguity (13.2%), 2.2% alter an operand and 1.9% have a well-posed question whose answer differs from the stored gold (Table 11). Validity is flat across languages (81.5–83.8%) and falls with complexity (88% at the lowest level to 75% at the highest). On the valid core every model gains 2–5 points and the ranking is *identical*; conditioning the interventions of §6 on it moves pooled effects by at most 1.2 points, with the large ones growing (Table 12), and no single model's by more than 3.1, and shrinks no language spread (Table 13). (ii) Mutation testing shows why a numeral gate alone is not enough: it catches 100% of edited or deleted operands and 0% of swapped operands, flipped relations and changed units, which model readers catch 80–93% of the time (Table 15); the per-item census labels supply that second layer for the test split.

Measurement properties. Across seven models the three splits agree within 1.4 points, every public-minus-private interval contains zero, and the splits rank the models identically (Table 20). With the text reduced to its numbers models score 0.03, and with words shuffled 0.02–0.12: the text must be read. Re-weighting languages to a common domain and tier mix moves accuracy by at most 0.30 points, so language gaps are not composition artefacts.

5 What the benchmark separates

Language spread is the headline, not accuracy. Table 3 and Figure 4 report seven open-weight models zero-shot. At matched decoding (reasoning off) best-minus-worst language spreads range from 11.0 (Sarvam-105B) to 59.6 points (Nemotron), and from 2.2 to 86.2 as served; consistency is not a function of accuracy: Sarvam-105B outscores Gemma-4-26B overall yet disagrees with itself on lineage-matched pairs half again as often (18.6% against 12.0%). A differential-item-functioning



Figure 4: Where the models fail: zero-shot accuracy by language with each model decoded as served (matched-decoding spreads are in Table 3), seven models, 88,030 scored items (6,786–7,861 per language). The right-hand column is each row's best-minus-worst language spread. A model that is near-ceiling in Hindi can lose half its accuracy in Odia or Malayalam; aggregate accuracy hides this.

Table 3: Seven open-weight models on BharatMath. *Left:* full-corpus zero-shot campaign (88,030 scored items, each model as served; the two Gemma-4 rows produced no reasoning trace and are effectively thinking-off). Spread is best-minus-worst language accuracy in points; flip rate is the share of 13,811 lineage- and answer-matched cross-version pairs solved in only one version. The left block mixes reasoning states by design (it is what a user of each served model gets) and is not used to rank; every comparison in this paper uses the right block. *Right:* the controlled campaign on the 8,860-item public test split with one prompt, one grader and the reasoning switch set explicitly; reasoning on is shown greedy and with the vendor's sampling (five samples; T=0.6 for Qwen3.8). ^{s}sampled decoding; $^{\ell}$greedy long-form decoding loops and truncates (12% and 34% of generations), so these two cells measure a decoding failure, not the model (§6); ^{3k}3,000-item stratified subset.

	Full corpus, as served			Test split, controlled		
Model	**acc.**	**spread**	**flip %**	**off**	**on, greedy**	**on, sampled**
Qwen3.8-27B	**0.940**	**2.2**	**6.2**	0.828	**0.922**	**0.926**
Gemma-4-31B-it	0.871	12.7	8.1	**0.871**	0.817^{ℓ}	0.921
Sarvam-105B	0.859	13.3	18.6	0.865^{s}	0.875^{s}	0.873
Gemma-4-26B-A4B-it	0.852	13.8	12.0	0.843	0.624^{ℓ}	0.881
Gnani Evon 3.3	0.846	21.2	20.7	0.661	0.828	0.843
Sarvam-30B	0.818	30.2	26.0	0.823^{s}	0.832^{s}	0.820^{3k}
Nemotron-3-Nano-30B-A3B	0.603	86.2	53.8	0.262	0.589	0.591

analysis over 23,097 source problems realised in several languages attributes 75% of the language-gap variance to model-specific effects and 25% to a component common to all models (Table 22): the gaps describe the models far more than the items.

Against a translated benchmark. Figure 5 (left) runs the same models, prompt and grader on MGSM, and on BharatMath at MGSM's size. MGSM-English spans 0.936–0.988, within which a model that reads no Odia is indistinguishable from the strongest. MGSM-Bengali already separates these models (0.47–0.92); what changes on BharatMath is the order. BharatMath-Bengali is *harder* than translated GSM8K for the general-purpose models (Gemma-31B 0.916 → 0.784, Nemotron 0.472 → 0.276) but *not* for the Indic-specialised ones (Sarvam-105B 0.880 → 0.864, Sarvam-30B 0.844 → 0.852), so the ranking changes; Telugu re-orders them too (Qwen3.8 0.656 → 0.752, Sarvam-30B 0.848 → 0.696; Table 18). MGSM cannot leak into BharatMath: a holdout gate drops every GSM8K-test source. Regenerating 203 of MGSM's own problems with operands and gold held fixed (723 problem–language pairs), our rendering is harder than the literal translation for all seven models (Table 25): our items are not an easier restatement of their sources. The Sarvam models still lead on the 62 GSM8K-derived Bengali items (intervals overlap; Table 19). Size matters as well (Figure 5, right): at 250 items per language, a third of random draws invert at least one model pair.

A matched language-by-framing comparison. A matched 2×2 over 600 contracts crosses language (English, Indic) with framing (source problem, regional re-telling). Both Indic cells are held to the same number-preserving gate (the translator loses a number in 12.2% of contracts). In English, regional framing costs -0.020 pooled (n.s.; at most five points, significant for two models). In the Indic languages the released regional item is solved *more* often than a machine translation of the

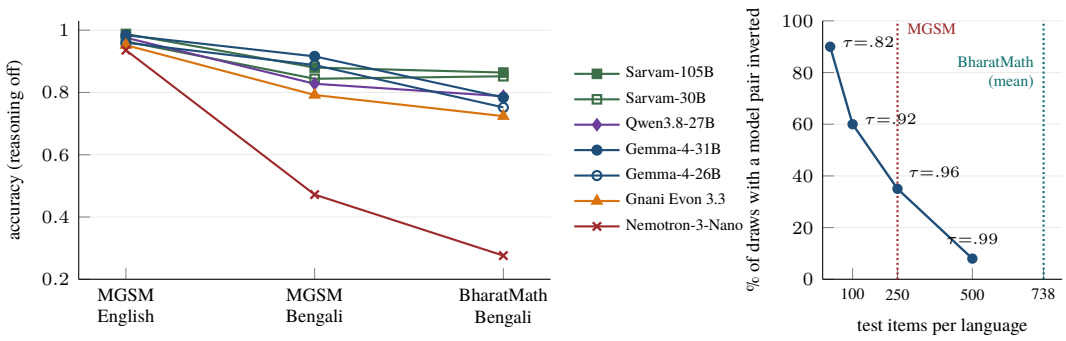


Figure 5: Translated against regenerated Bengali. *Left:* the same seven models, prompt and grader on three test sets. English is saturated (five points end to end). Translated Bengali spreads the models; regenerated Bengali re-orders them: the two Indic-specialised models move from third and fourth to first and second, and the general-purpose models lose 4–20 points (Spearman $\rho = 0.46$ between the two Bengali rankings, on the 250-item comparison of Table 18). *Right:* share of random subsamples in which the seven-model ranking has at least one inversion against the full test split (τ: mean Kendall correlation). At MGSM's 250 items per language a third of draws mis-rank a pair; BharatMath's public test split has 738 per language on average.

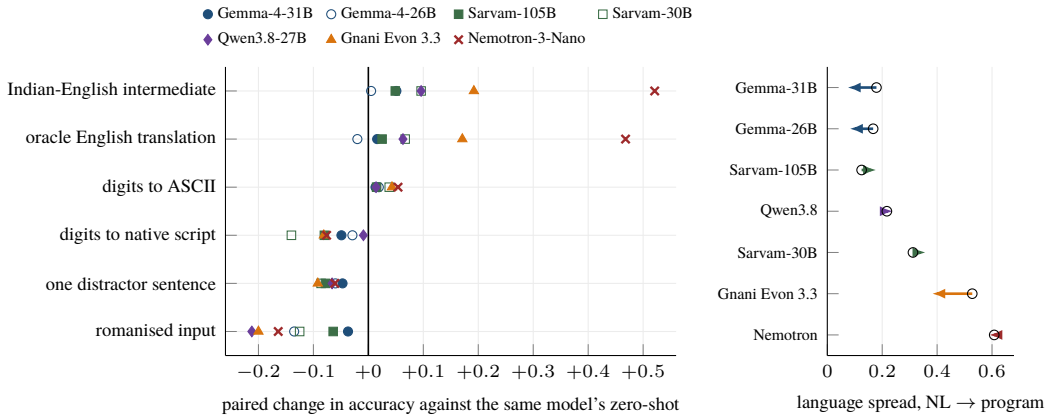


Figure 6: Single-factor interventions on identical items, reasoning off. *Left:* each mark is one model's paired change in accuracy when one thing about the input changes and the contract does not (full split, or the 3,000-item subset for the oracle-English, distractor and romanised arms; Sarvam models sampled). Reading the question in English is worth nothing to one model and 52 points to another; native digits hurt every model; one irrelevant sentence costs 5–9 points everywhere. *Right:* best-minus-worst language spread when the model answers in natural language (circle) and when it writes a program that code executes (arrow head). Removing arithmetic halves the spread of both Gemma models and of no other family. Tables 23–28 give every cell with intervals.

source: +0.038 [+0.018, +0.058] pooled, significant for five of seven models and never significantly lower (Table 25). On number-preserving contracts, regional framing does not make an Indic item harder than a translation of its source.

6 Six evaluation gaps

Each intervention changes one thing at a time. All contrasts are paired on identical items; intervals are language-stratified item bootstraps with 10,000 draws (protocol and every arm in Appendix E).

Gap 1: understanding the question. Every item retains its Indian-English intermediate: the same contract, numbers and Indian scenario, in English. Solving that instead of the final question is worth +0.005 (n.s.) to Gemma-26B and +0.05 to Gemma-31B, but +0.19 to Gnani Evon 3.3 and +0.52 to Nemotron, for which the language of the question is two thirds of all error. The two Indic-specialised Sarvam models still gain 5–10 points. An oracle translation of the final question gives almost the same gains, and merely asking a model to translate first and then solve recovers +0.17 (Nemotron) and +0.15 (Gnani Evon 3.3): the weak models *can* translate but do not unprompted.

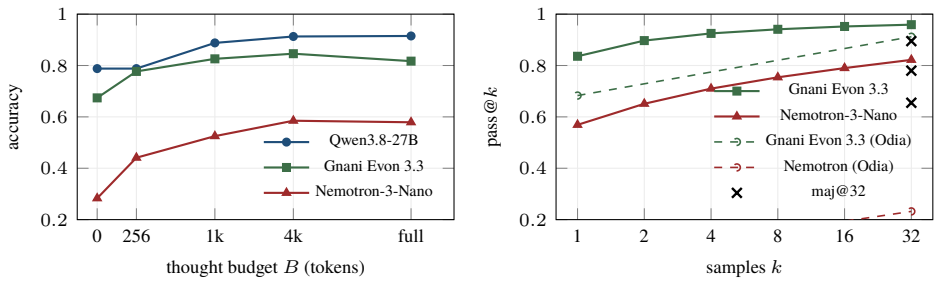


Figure 7: *Left:* budget forcing on 3,000 items. The model's own thought is cut at B tokens and an answer is forced. Most of the value of reasoning arrives within 1,024 tokens and a 4,096-token budget loses nothing; where it *beats* the unbounded thought, the cause is greedy looping. *Right:* pass@k from 32 samples. Crosses mark majority vote over the same 32 samples, the deployable counterpart of the pass@32 oracle (0.895 and 0.655 overall, 0.78 for Gnani Evon 3.3 in Odia): about half the headroom is reachable by voting. For Gnani Evon 3.3 the Odia answer is within 32 samples for 91% of items; for Nemotron, 23% (the dashed Nemotron line leaves the plotted range at k=1, 0.04).

Gap 2: numeral script. Rewriting only the digits of the question to ASCII helps every model ($+0.013$ to $+0.054$) and rewriting them to the native script hurts every model (-0.009 to -0.140); the largest penalty belongs to an India-built model, Sarvam-30B. Byte-identical twins of 6,000 renderings that differ *only* in digit script make this causal: native digits cost 0.02 (Qwen3.8) to 0.34 (Nemotron, reasoning on), with the predicted sign and an interval excluding zero in all sixteen model-by-direction cells (Table 27). Requiring a native-numeral *answer* costs a further 2.5–21 points.

Gap 3: an execution component, for one family. Asking for a straight-line program, executed by a restricted interpreter, removes arithmetic from the task (Figure 6, right). For the two Gemma models this *halves* the language spread (0.179 $\rightarrow$ 0.072, 0.167 $\rightarrow$ 0.080), with the gain in the two languages flagged as Gemma-specific weaknesses. Of Gemma-31B's natural-language failures, the program is correct for 73% in Assamese and 72% in Bengali (58% in Hindi; Table 29): consistent with a deficit in carrying out the computation in natural language, though a program prompt may also change how the question is read. For the other families the spread survives code execution (Nemotron 0.61 $\rightarrow$ 0.59, Sarvam-30B 0.31 $\rightarrow$ 0.36, Qwen3.8 0.22 $\rightarrow$ 0.24) or narrows only together with accuracy (Gnani Evon 3.3, 0.53 $\rightarrow$ 0.38 at -0.059): for these four the program is right on only 15–41% of failures (Gemma-31B: 61%), which points upstream of execution. With the program register held fixed, English input adds nothing for Gemma-31B (-0.011; a calculator alone gives $+0.034$) but lifts Gnani Evon 3.3 by $+0.126$ and Nemotron by $+0.415$ (Table 30): consistent with an execution component for Gemma-31B and a comprehension component for Gnani Evon 3.3 and Nemotron; a fixed output register reduces, but does not exclude, code competence as a contributor. Scale does not close it either: across four Gemma-4 sizes accuracy climbs 0.685 $\rightarrow$ 0.871 while the spread stays at 0.13–0.20, with Assamese and Bengali the floor at every rung (Table 31).

Gap 4: sampling-accessible headroom. For Gnani Evon 3.3, pass@1 of 0.836 rises to pass@32 of 0.959, and the headroom is largest exactly where pass@1 is lowest: Odia 0.683 $\rightarrow$ 0.912, Malayalam 0.772 $\rightarrow$ 0.944 (Figure 7, right). Sarvam-30B (0.805 $\rightarrow$ 0.950; lowest language 0.84) and Sarvam-105B (0.852 $\rightarrow$ 0.969) behave the same way. An oracle over samples is not a deployable gain: majority vote recovers about half of it (maj@32 0.895, 0.876 and 0.908 for the same three models; Odia 0.78 for Gnani Evon 3.3). Reasoning itself is worth $+0.09$ (Qwen3.8), $+0.17$ (Gnani Evon 3.3) and $+0.33$ (Nemotron), most of it within 1,024 thought tokens (Figure 7, left). Greedy decoding with reasoning on is not a conservative estimator: Gemma-4-26B loops on 34% of items and scores 0.624, *below* its reasoning-off 0.843 and its sampled 0.881; truncation shares for the reasoning-on arms are in Tables 36, 33 and 34.

Gap 5: the cost of forced native-script thought under a fixed budget. Left free, models think in English (native script is 0.7–15.3% of free thought), whatever the instruction. Prefilling one target-language sentence at the start of the thought makes 93–99% of thoughts native, and against a prefilled *English* sentence as control accuracy falls by 0.204 for Gnani Evon 3.3, 0.178 for Nemotron and 0.040 for Qwen3.8 (all intervals exclude zero) at 1.3–2.4$\times$ the tokens (Table 33). Most of that coincides with the token budget: native thoughts hit the cap on 37% of items for Gnani Evon 3.3 against 14%,

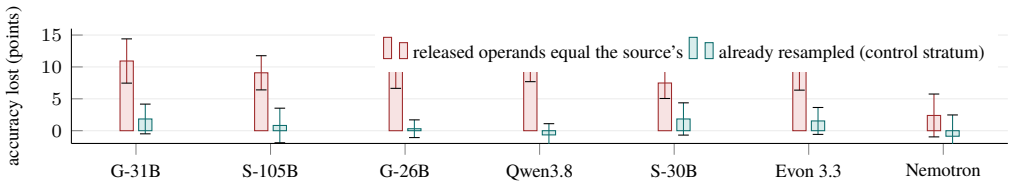


Figure 8: Accuracy lost when operands are re-minted at the same magnitude, text kept (1280 contracts, three draws, reasoning off; 95% language-clustered intervals). Already-resampled items get the identical procedure and show no detectable average drop (-0.7 points, interval includes zero); the measured loss is concentrated on items whose operands coincide with the public source. The strata are not randomly assigned (Tables 38, 39).

and on items both arms complete the loss is 0.042 (Gnani Evon 3.3), 0.011 (Qwen3.8) and 0.165 (Nemotron) (Table 34). On completed items only Nemotron loses much.

Gap 6: robustness of reading. One irrelevant numeric sentence costs every model 5–9 points, and still 3–9 with reasoning on: reasoning does not filter a distractor. Romanised input costs 4–21 points, while paraphrase is nearly free (-0.027 to $+0.028$; -0.001 pooled on the valid core, n.s.), so items are not brittle to wording (Table 36). In-language few-shot exemplars help the Gemma models and *hurt* Qwen3.8 (-0.044), Sarvam-105B (-0.092) and Sarvam-30B (-0.024; Table 37), so k-shot evaluation understates three of seven models. And re-rendering one contract changes the verdict for 35–65% of contracts, so a single rendering per problem confounds the item with its wording.

7 Sensitivity to source-identical operands

BharatMath inherits problem *structures* from public sets, so we ask whether models recognise them. Recitation is rare: a wrong answer equals the source problem's gold in at most 0.2% of responses. Re-minting measures the rest, and the corpus supplies its own control stratum. Stage 1 resamples every operand unless Stage 0 emitted a degenerate range for it, so 30.6% of public test items (28.6% of the corpus) still carry their source's numbers and the rest were already resampled once. On all 1,280 eligible contracts we resample operands at the same magnitude, keeping the text, three times, for seven models. The pooled drop is -0.045 [-0.056, -0.036], and it is concentrated on the 628 source-identical items: -0.085 [-0.102, -0.068] there against -0.007 [-0.018, $+0.004$] on the already-resampled items, a difference of -0.078 [-0.100, -0.058] that holds for six of seven models (Figure 8, Table 38). The strata were not randomly assigned, so differential changes in validity or difficulty remain possible: on pairs a judge calls clean in both versions the source-identical loss is 3.6 points [0.3, 6.9] and its difference from the control is not resolved (Table 39). Integer golds (-0.070), rounder numbers and arithmetic load do not remove it (Appendix F). The loss is what source familiarity predicts; that source authors chose friendlier numbers than a sampler does is not excluded, and we claim a sensitivity to source-identical operands, not a contamination rate. It matters for the leaderboard: on already-resampled items the two Gemma models score ten points lower than on source-identical ones and the Sarvam models under five, so the top five compress: the three tiers hold in both strata, the order inside the leading tier does not (Gemma-31B 0.941, Sarvam-105B 0.895 source-identical; 0.840, 0.852 resampled). We release the flag with every item and recommend reporting the resampled stratum.

8 Limitations

The defects below are measured or disclosed, not repaired. **Validity is measured by models.** Every validity and cultural-quality figure comes from LLM readers and judges ($\kappa = 0.42$ between families); no native speaker has rated the items; 17% of labelled test items are not fully valid, and the valid core is model-adjudicated, not human-validated; the ranking is the same on it. **The bank is bounded.** The evaluated build uses two of its four fields; 4.4% of its records are disputed by both auditing families, and items carry no record identifier: which record an item used is recovered from the lineage for about half of them. **Artefacts, flagged per item:** 58% of questions print an integer operand as a float, and 7.3% of formulas are insensitive to a variable. **Scope.** Seven open-weight models and a four-rung scale ladder, no closed, mathematics-specialised or long-chain-of-thought model; studies that train on the release should report the 2,270 test items with no training-side near-twin; the interventions are paired but not mechanistic, and perturbations are machine-made.

AI use statement

Generative AI is part of the object of study: the benchmark's questions and solutions were generated by open-weight language models inside the pipeline of §3, bank cells were drafted with a frontier model and reconciled against public reference data by the authors, and language models serve as blinded raters, readers and judges in §4–7, each with a reported error rate. We also used generative AI tools for writing analysis code, for orchestrating and monitoring the evaluation campaign, and for drafting and editing the text of this paper. We did not use generative AI to originate the research idea. All AI-assisted code was tested, every number in the paper is regenerated by released scripts from released predictions, and the authors reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

Personas are synthetic records from a public synthetic dataset; no personal data were collected. Items name real districts, occupations and prices; regional portrayal was checked by pattern detectors and an LLM rater, not yet by native speakers from each region, and may contain stereotyped or outdated depictions. The corpus derives from publicly licensed English sources whose licences are respected in the release. The private test split is withheld to preserve its value and will be scored through a submission harness that returns accuracy only.

Reproducibility statement

The supplementary material contains the corpus with per-item lineage and quality flags, the three Context Bank workbooks, all prompts, the generation and evaluation harness, per-item predictions for every arm reported, and the analysis scripts that regenerate every table. Appendix B specifies the pipeline, Appendix E the evaluation protocol (models, decoding, token caps, grader, bootstrap) and a log of failed runs and their fixes. All generation, evaluation and auditing used open-weight models on self-hosted hardware with fixed seeds; the one exception is the blind rater of the four-method study, a proprietary model whose ratings we release.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of NAACL-HLT*, 2019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Israel Abebe Azime, Tadesse Destaw Belay, Dietrich Klakow, Philipp Slusallek, and Anshuman Chhabra. Bridging the culture gap: A framework for LLM-driven socio-cultural localization of math word problems in low-resource languages. In *Findings of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2508.14913.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*, 2023a.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dominika Agnieszka Długosz, Arlindo Oliveira, and Natalia Díaz-Rodríguez. The importance of being statistically earnest: A critical re-evaluation of GSM-Symbolic. *arXiv preprint arXiv:2605.28700*, 2026.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *International Conference on Machine Learning*, 2023.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1318–1326, Portland, Oregon, USA, 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1132`.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023.

Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. MMATH: A multilingual benchmark for mathematical reasoning. *arXiv preprint arXiv:2505.19126*, 2025.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

NVIDIA. Nemotron-Personas-India. Hugging Face Datasets, CC BY 4.0, 2025. 3M persona records with 21M persona descriptions, grounded in distributions from the 2011 Census of India and Indian labour statistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of NAACL-HLT*, 2021.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, 2023.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.19799.

Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. DIWALI – diversity and inclusivity aWare cuLture specific items for India: Dataset and assessment of LLMs for cultural text adaptation in Indian context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 33599–33626, Suzhou, China, 2025. Association for Computational Linguistics.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. DOSA: A dataset of social artifacts from different Indian geographical subcultures. *Proceedings of LREC-COLING*, 2024.

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. *arXiv preprint arXiv:2210.03057*, 2022.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Aditya Tomar, Nihar Ranjan Sahoo, Ashish Mittal, Rudra Murthy, and Pushpak Bhattacharyya. Mathematics isn't culture-free: Probing cultural gaps via entity and scenario perturbations. *arXiv preprint arXiv:2507.00883*, 2025.

Sang Truong, Serena Wang, Angelina Wang, Nhi Truong, Anka Reuel, Nick Haber, and Sanmi Koyejo. Public AI benchmarks are broken, but are private benchmarks the answer? Preprint, Stanford AIMS Lab, 2026. URL `https://aimslab.stanford.edu/papers/Truong2026PublicPrivate.pdf`.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2411.02538.

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

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, and Jingren Zhou. PolyMath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*, 2025.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. Epistemic diversity and knowledge collapse in large language models. *arXiv preprint arXiv:2510.04226*, 2025.

Fan Wu, Emily Black, and Varun Chandrasekaran. Generative monoculture in large language models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Minghao Wu, Weixuan Yuan, Longyue Wang, et al. The bitter lesson of multilingual benchmarks: A survey of over 2,000 multilingual benchmarks. *arXiv preprint arXiv:2504.15521*, 2025b.

Tianyi Xu, Kosei Uemura, Alfred Malengo Kondoro, Tadesse Destaw Belay, Catherine Nana Nyaah Essuman, Ifeoma Okoh, Ganiyat Afolabi, Ayodele Awokoya, and David Ifeoluwa Adelani. MGSM-Pro: A simple strategy for robust multilingual mathematical reasoning evaluation. *arXiv preprint arXiv:2601.21225*, 2026.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, et al. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*, 2024.

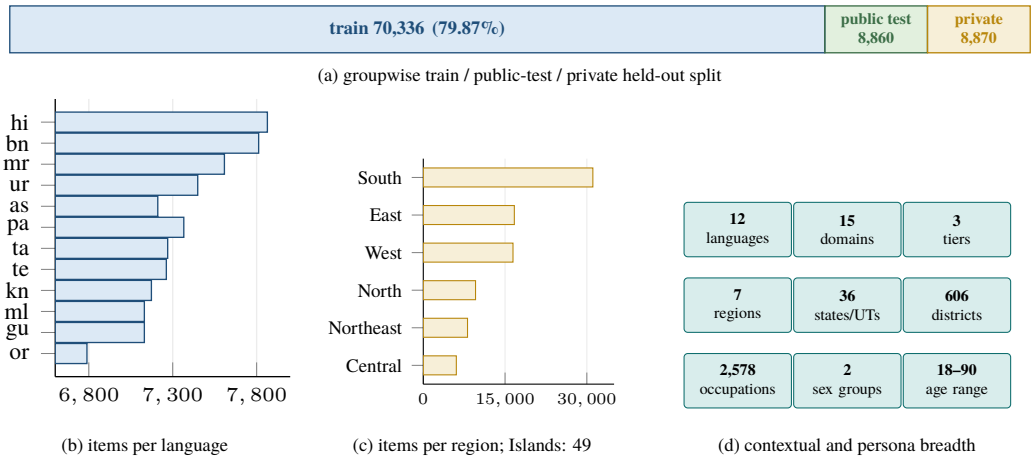
Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*, 2025.

A The benchmark at a glance

Table 4 lists the three layers of the Context Bank, Figure 9 the composition of the corpus and its splits, and Table 5 one bank record per language. The last is the shortest answer to "what makes this Indic": the units a Tamil land problem may use are not those of a Punjabi one, and the bank says so explicitly.

Table 4: The three layers of the 782-cell BharatMath Context Bank. Entries describe the bank specification; Appendix B distinguishes it from the released routing.

Layer	Returned context	Cells	Example
P1: units	u: allowed and forbidden units	438	HI/LAND: *bigha, biswa, katha*; region-specific *bigha* conversions.
P2: world	ρ, σ: ranges, rates and scenarios	259	WAGE LABOUR/BHATTA: local piece rates, daily output and seasonal duration.
P3: language	λ: language-structural guidance	85	Hindi verb-final exemplar and an explicit SVO failure pattern.



(a) groupwise train / public-test / private held-out split

(b) items per language

(c) items per region; Islands: 49

(d) contextual and persona breadth

Figure 9: Corpus composition and split. Language totals are near-uniform (6,789–7,864), while contexts cover every major Indian region and diverse persona profiles. The public test and private held-out sets each cover 569 districts. Each language contributes 685–792 public-test items, and the largest axis-level distribution gap between the two sets is TVD 0.0025. Table 20 compares model performance on these partitions. Domain frequencies follow the source distribution rather than a uniform allocation.

Language-specific records. Pass 1 holds 438 unit cells, one per (L, D) key; Pass 2 holds 259 world records (223 language-tagged scenario rows, 12 pan-India reference-rate blocks and 24 scenario-map blocks); Pass 3 holds 85 (12 native-language re-voicing preambles, 37 structural exemplars and 36 translationese rules). 425 of the 438 unit cells (97.0%) carry an explicit forbidden list, and 17 units are *contested* inside a single domain: preferred in at least one language and forbidden in at least one other. *Bigha* is a preferred unit in land and agriculture cells of seven languages and names three different areas (27,225 sq ft in Purvanchal Hindi, 17,424 in western UP and as Gujarati *vigha*, 14,400 in Bengali and Assamese). Preferred and forbidden lists are properties of an (L, D) cell, not of a language, so the counts overlap: every language forbids *bigha* in its school-geometry cell, where textbooks use square metres, and Punjabi forbids it for land as well, keeping the name for its own four-*kanal* unit.

Table 5: Example records from the three Context Bank workbooks. The upper block is Indo-Aryan and the lower Dravidian; × marks units forbidden by the bank in the stated domain. The λ column records register-specific prompt heuristics, not validated linguistic generalisations; native-speaker review remains necessary. These are bank specifications, not guarantees about generated text. The Pass-3 preambles and exemplars were omitted from the released prompt; the routing of its rules is not specified here.

Lang.	u — units, and what is forbidden	ρ, σ — world and magnitudes	λ — language structure
hi	*bigha*, *biswa*, *seer* (933 g), *maund*; × acre, hectare	UP cane SAP Rs. 370/quintal; brick-kiln piece rate Rs. 450–650 per 1,000 bricks	ergative *ne*, verb agrees with object gender; *kitnā/-ī/-e* declines
ur	*tola* (11.66 g), *masha* (0.97 g), *ratti* (0.12 g); × sovereign, guntha	zakat: *nisab* 85 g gold or 595 g silver, rate 2.5%	Perso-Arabic maths lexicon (*zarb*, *taqsim*, *sood*); "per kg" is *fi kilo*
pa	*kanal*, *marla* (30.25 sq yd), *ghumao* = 8 kanal = 1 acre, *murabba* (25 acres); × bigha (UP variant), katha	*arhatiya* commission 2.5% + Rs. 2.50/quintal; stubble fine Rs. 2,500/acre	ergative *ne*; *kinnā/-ī/-e* declines by gender–number
gu	*guntha* (1,089 sq ft = 1/40 acre), *vigha*; × kanal, marla, katha	Surat diamond polishing Rs. 2–20/stone; cotton bale 170 kg, ginning Rs. 800–1,200	oblique before postposition; the calque *6 keri Rs. 90 ni kimat dharave chhe* is rejected
mr	*guntha* (40 = 1 acre), hectare in land records; × bigha, kanal, marla	Lasalgaon onion Rs. 500–3,000/quintal, *adatya* 4%; cane FRP Rs. 340/quintal	postposition fuses with the oblique (*dukan*+*at* → *dukanat*); the Hindi system is forbidden
bn	*bigha* (14,400 sq ft = 20 *katha*), *bhori* (11.66 g gold); × guntha, kanal, sovereign	jute MSP Rs. 5,335/quintal at 10–16 *maund*/*bigha*	no grammatical gender; definiteness suffix *-ti/-ta*; *kiloprati*, not a slash
or	*kani* (≈1/3 acre), *decimal*, *guntha*; × bigha, katha, marla	Mayurbhanj sal-leaf bundle Rs. 20–60; mahua 10–50 kg/week at Rs. 20–80/kg	the rupee is *tanka*, never a transliteration; "per kg" spelled out
as	*bigha* (14,400 sq ft = 5 *katha*), *lecha* (144 sq ft); × decimal (a WB term), guntha	tea 2,000–2,500 kg made tea/ha/yr; plucking wage Rs. 250–350/day	directional *-loi* against locative *-ot*, a named model error
kn	*guntha*, *cent* (south Karnataka), *ankanam* (72 sq ft); × bigha, kanal, marla	silk cocoon Rs. 350–550/kg; 1 kg cocoon → 80–100 g raw silk	agglutinative case on the stem; verb agrees person–gender–number
te	*guntha*, *ankanam* (36 sq ft, urban Telangana), *puti* (≈25 kg); × sovereign, maund	Guntur chilli 50–80k bags/day, Teja Rs. 8,000–20,000/quintal	*enta* (amount) against *enni* (count); "per unit" is *-ki/-ku*
ta	*cent*, *ground* (2,400 sq ft), *kuli* (72 sq ft), *padi* (≈1.25 kg); × bigha, guntha, *para*	Amma canteen *idli* Rs. 1, sambar rice Rs. 5; Kanchipuram weaver Rs. 300–600/day	formal *ezhuttu* register; the wh-word never stands first
ml	*cent* (1/100 acre), *pavan*/sovereign (8 g), *para* (8–13 kg), *nazhi* (200 g); × tola, bigha, guntha	RSS-4 rubber Rs. 170–210/kg; KSFE *chitty*, foreman commission 5%	most agglutinative of the twelve; participial adjective precedes its noun

B The construction in full

Figure 10 follows one item through the pipeline and Table 6 states the procedure. Stage 0 is a model extraction followed by a deterministic check, and Stage 1 is deterministic code with no model call; together they construct the read-only contract. Stages 2 and 3 generate question and solution text, subject to the listed checks.

What the released prompts carry. Of the four bank fields, three reach the generator. Units u enter both question calls, the preferred list as mandatory and the avoid-list as a hard constraint; scenarios σ enter Call A as ranked options with cultural notes; magnitudes ρ enter Call A as a reference-rate block and per-scenario hints. Language guidance λ is retrieved, and a job missing it is rejected, but it is never assembled into a prompt. The design also calls for intersecting ρ with blueprint ranges at sampling time; in the released code a key mismatch meant that none of the $353{,}096$ variable slots matched a bank constraint, so all values came from blueprint ranges alone. The bank effect measured in Appendix C is therefore an effect of u, σ and advisory ρ. Unit roles are free text, not typed quantities, so a substitution such as *acre* to *bigha* changes the stated unit but not V' or y', and dimensional consistency is not checked in code; this is one reason we audit question validity separately (Appendix D).

Rejection, not repair. Jobs that exhaust a retry budget or fail a final check are discarded: a blueprint whose sampled valuation cannot meet the plausibility checks within 20 draws, an exhausted bank backoff chain, a missing preamble or exemplar, a solution whose answer cannot be extracted or misses

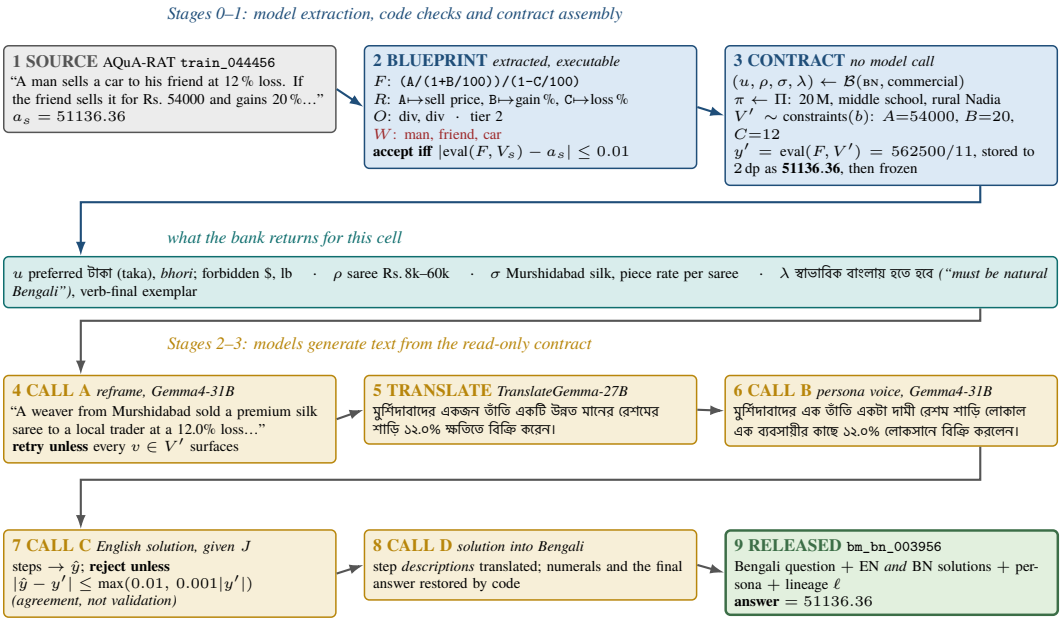


Figure 10: One item through the pipeline. J is the Stage-1 job record, which carries F, V', y', the bank record and the persona. A model extracts the formula; code checks its value against the source answer; the bank record and persona are bound *first*, and V' is then resampled under the blueprint's own numeric constraints, after which y' is frozen. The bank's magnitude ranges ρ do not bind the draw in the released build, and its language guidance λ is stored and validated but never reaches a prompt; §B says why. Later calls may rewrite the setting and wording but cannot alter F, V' or y'. What each gate does and does not certify is stated below Table 6.

Table 6: The pipeline as routing and gates. **The gold y' is fixed by code at Stage 1, before any question text exists, and no call in Stages 2–3 has a write path to F, V' or y'** — Stages 2 and 3 write only text, subject to the specific checks listed above. Stage 0 is the only model call that touches the invariant, and a deterministic check follows it immediately. Figure 10 traces one item through all four stages. Prompt contents, retry budgets, persona eligibility and the bank backoff chain $(L, D) \to (L, D') \to (L, *) \to (*, D)$ are in the release documentation.

Stage	Who acts	What it may write	The gate rejects the item if
0 blueprint extraction	Gemma4-31B extracts; code checks	F, V_s, R, O, S, W	$\lvert\text{eval}(F, V_s) - a_s\rvert > \varepsilon_{\text{ext}}$
1 contract assembly	**code only** (no model call)	V', $\mathbf{y}'$, $(u, \rho, \sigma, \lambda)$ from $\mathcal{B}$, π, ℓ	the bank backoff chain exhausts; y' is undefined, non-finite or implausible after 20 draws; the re-check $\lvert\text{eval}(F, V') - y'\rvert > \varepsilon_{\text{ext}}$ fails, or a field of J is absent
2 question realisation	Gemma4-31B → TranslateGemma-27B → Gemma4-31B	the question text only (t_{en}, t_{mt}, t_L)	a sampled value of V' is not surfaced in t_{en} (at most two retries). *No further check is applied to the question after this point*
3 solution and export	Gemma4-31B; code restores numerals	the solution text only (s_{en}, s_L)	no answer can be extracted; $\lvert\hat{y} - y'\rvert > \max(0.01,\ 0.001\lvert y'\rvert)$; the row is script-impure, contaminated, or its lineage is broken

the gate. No failing job is edited in place; the one post-hoc intervention, for 5,490 rows with a script defect in the question text, re-ran Stages 2 and 3 against the same frozen contract. Rejection rates differ across cells, so against a uniform twelve-language target the release runs from 8.93% of rows (Hindi) to 7.71% (Odia). On the 82,576 rows whose archived traces record both the generated answer and the gold, 99.93% agree with y' at the precision written, and 56 rows (0.068%) pass only through the relative tolerance of the Stage-3 gate; they are listed in the release.

C Validating the Context Bank

C.1 Four generation methods on shared contracts

Conditions. C1 translates the English source, with sampled values substituted, using one `indictrans2-en-indic-1B` call. C3 generates directly in the target language with one `DeepSeek-V4-Flash` call. C2 and C4 use the release's three question calls (`gemma-4-31B-it`, `translategemma-27b-it`); C2 replaces the bank blocks with "Adapt this problem for an Indian setting" and keeps persona, register, blueprint, few-shot examples and output contract. C4 is taken from the released corpus rather than generated anew, so C4−C2 compares the same configured pipeline with and without banks but with different selection histories; we therefore report a gate-matched subset next to the full sample. Restricting both arms to first attempts (599 C4 and 541 C2 items) gives $+11.4$ points [$+7.3$, $+15.2$] of regional appropriateness against $+11.3$ [$+7.7$, $+15.0$] on all rated items, so retry history does not produce the contrast. C4−C3 and C4−C1 also vary the model and call structure and are end-to-end comparisons.

Sampling and scoring. 50 items per language, stratified by tier and operation type, frozen before C1–C3 were generated. Candidates are shuffled into unlabelled slots within each item and scored 1–5 against prespecified anchors; 4 or 5 passes. Paired per-item differences are bootstrapped over the 12 language clusters ($B = 2{,}000$) and Holm-corrected across the three reliable axes. Repeatability on a 10% re-rating sample is $\kappa = 0.916$ (mathematical equivalence), 0.875 (regional appropriateness) and 0.774 (cultural plausibility); a fourth axis, provenance correctness, reached 0.339 and is excluded.

Table 7: Paired generation comparison on 600 contracts, 50 per language. Formula, sampled values and gold answer are fixed across conditions; only the question is rated. Pass rates use the full sample, while contrasts below the rule use the gate-matched subset. C4 reuses released outputs; C1–C3 are newly generated. Scoring uses a single LLM rater with shuffled condition labels, although ten items were partially unblinded. The text describes the design, selection differences and sensitivity analyses.

Arm	components present				pass rate (score ≥ 4 on 1–5)		
	literal MT of source	generative rewrite	3 question calls + persona	Context Bank	math. equiv.	cultural plaus.	regional approp.
C1	■	□	□	□	0.702	0.007	**0.000**
C3	□	■	□	□	0.843	0.118	0.228
C2	□	■	■	□	**0.938**	0.498	0.530
C4	□	■	■	■	0.878	**0.517**	**0.643**
C4 − C2 *(bank-on vs. bank-off)*, gate-matched					-0.053	$+0.013$	$+$**0.091**
p (Holm), paired language-clustered bootstrap					0.010	*0.579*	≤ 0.002

Limitations of the rater. One LLM, not a native-speaker panel, rates all questions. The rater is a Claude model, a different family from both generators, which reduces self-rating; it is, however, the same agent that built and ran the pipeline, so unbiased judgement is not established, and on ten items (1.7%) it had seen raw C2 output during debugging. These items remain in the analysis. C4's 73 non-passing mathematical-equivalence items split $16/28/29$ across scores $1/2/3$: 2.7% pose a different problem or drop a frozen value, and 4.8% stay answerable at the stored gold but not uniquely so. The 12.2% complement of the pass rate is therefore not a wrong-answer rate.

C.2 Twelve arms on fresh contracts

Design. We draw 600 *fresh* contracts (50 per language, distinct source problems): the released blueprint with newly sampled operands, every operand different from the released draw, so that no arm can reproduce a released question. Gemma-4-31B-it generates at $T = 0.7$, three seeds, *without retries* and without few-shot exemplars in any arm; the study therefore measures first-attempt behaviour and is a lower bound on shipped yield. The arms differ only in the conditioning blocks shown to the generator: `full`; `bank_only`; `persona_only`; `none`; `alloc_only` (state and district named, nothing else); four leave-one-layer-out arms; `no_district`; `miskeyed` (units, scenarios and rates taken from a contract of another language); and `direct`, which writes the target language in one call. Outcomes are computed per attempted contract with no filtering: every frozen operand present and the

gold absent (code); a blind Qwen3-235B reader, shown only the question, recovers the fresh gold; and the state a blinded Qwen3-235B judge extracts from the text equals the persona's state. Gemma-4-31B repeats the judge call as an agreement check. All 21,600 generations were read and judged.

Table 8: Twelve-arm generation study: 600 fresh contracts × 3 seeds per arm, first attempt, no filtering. Y_{num}: operands present and gold absent; Y_{valid}: also recovered by the blind reader; Y_{reg}: the blinded judge extracts the persona's state; Y_{joint}: both; "district": the persona's district is named in the text. Ratings are 1–5 means. "scen." is the number of distinct Call A scenarios among the 1,800 attempts.

arm	Y_{num}	Y_{valid}	Y_{reg}	Y_{joint}	district	specific.	natural.	plausib.	scen.
`full`	.973	.637	.279	.175	.089	2.38	4.87	4.86	649
`bank_only`	.976	.621	.442	.292	.009	3.31	4.93	4.86	315
`persona_only`	.959	.618	.092	.057	.049	1.59	4.79	4.88	982
`alloc_only`	.942	.582	.966	.565	.750	4.47	4.98	4.88	1116
`none`	.937	.603	.062	.039	.002	1.41	4.93	4.91	663
`no_u` (units)	.976	.627	.292	.177	.084	2.47	4.85	4.85	599
`no_s` (scenarios)	.964	.619	.167	.104	.098	1.84	4.82	4.88	919
`no_r` (rates)	.973	.636	.287	.182	.089	2.45	4.87	4.87	632
`no_l` (linguistic)	.975	.637	.279	.177	.088	2.37	4.87	4.88	649
`no_district`	.973	.624	.269	.165	.003	2.41	4.86	4.87	603
`miskeyed`	.981	.631	.169	.111	.094	2.04	4.84	4.88	884
`direct`	.984	.641	.170	.103	.009	1.94	4.89	4.83	497

Reading the table. Validity does not differ detectably across arms with this generator; that is an absence of a detected cost, not an equivalence result, and it does not transfer (with Qwen3-235B generating, the bank contrasts lose 12.6 and 7.3 points of validity; Table 9). Regional recoverability is carried by the scenario layer: removing it, or keying the bank to the wrong region, returns the arm to the level of a single direct call. The two results that cut against the shipped design are visible in the first two rows: adding the persona to the bank doubles the number of distinct scenarios and lowers the recoverable-state rate, and `alloc_only`, which simply names the place, tops the table. The regional outcome rewards naming; the four-arm study above, which scores appropriateness rather than recoverability, is the complement.

A second generator. We replay the study with Qwen3-235B as generator, everything else identical; because Qwen then judges its own generations, Gemma-4-31B repeats every judgement out of family (agreement on the state 92.0%, $\kappa = 0.83$). Every grounding contrast replicates in sign and approximate size (Table 9). Fidelity does *not* transfer for free: on its first attempt Qwen drops at least one frozen operand from 34.8% of questions, against 3.2% for Gemma. The shipped numeral gate rejects every such draft and retries, so this bounds the *cost* of switching generator, not the quality of what is released.

C.3 Can a judge tell grounding from name-dropping?

State recoverability rewards naming a place, and a cultural-quality measure should not. We build seven controlled versions of 600 public-test items that differ only in grounding: *released*; *neutral* (every regional detail removed); *names* (local personal names only); *places* (named places only); *noplaces*; *explained* (one clause explaining the local unit); and *mismatched* (moved to a distant state). Qwen3-235B and Gemma-4-31B score each text blind to its variant. The blind reader confirms the variants are the same problem (validity 0.781–0.805, no interval against the released text excludes zero). Names do almost nothing (names beat neutral on 23–30% of items); places alone recover all but 0.1–0.4 points of the released specificity; moving a problem to the wrong state costs 1.6–2.3 points of regional fit and 0.1–0.2 of plausibility (Table 10). Qwen agrees with itself across batches ($\kappa = 0.976$); across families exact agreement is 48% (weighted $\kappa = 0.73$) with every contrast of the same sign, and pairwise preference favours the longer text 79–81% of the time, so we report absolute scores. The caveat that bounds our cultural claim: only 41–58% of released problems are judged *substantively* grounded as opposed to place-only, and no native speaker has yet calibrated this judge.

Table 9: The twelve-arm study with two generators (600 contracts × 3 seeds per arm, first attempts, no retries). Operands: every frozen operand appears in the question. Regional: the blind judge recovers the contract's state; the out-of-family judge's value is in parentheses. Paired contrasts use 4,000 bootstrap resamples of contracts within language.

	Gemma-4-31B generates			**Qwen3-235B generates**		
Arm	Operands	Valid	Regional	Operands	Valid	Regional
none (no persona, no bank)	0.937	0.603	0.062 (0.049)	0.783	0.472	0.029 (0.022)
persona only	0.959	0.618	0.092 (0.118)	0.666	0.398	0.137 (0.189)
bank only	0.976	0.621	0.442 (0.480)	0.704	0.346	0.464 (0.509)
full (shipped)	0.973	0.637	0.279 (0.331)	0.598	0.325	0.436 (0.487)
full, scenarios removed	0.964	0.619	0.167 (0.198)	0.648	0.368	0.229 (0.304)
full, bank mis-keyed	0.981	0.631	0.169 (0.199)	0.582	0.327	0.238 (0.298)
direct (no pivot)	0.984	0.641	0.170 (0.203)	0.752	0.406	0.171 (0.219)
allocation named only	0.942	0.582	0.966 (0.971)	0.796	0.469	0.784 (0.788)
bank effect, no persona		+0.018	+0.381*		−0.126*	+0.435*
bank effect, with persona		+0.018	+0.187*		−0.073*	+0.299*
persona × bank interaction		+0.000	−0.194*		+0.053*	−0.136*
scenarios removed		−0.017	−0.112*		+0.043*	−0.207*
bank mis-keyed		−0.006	−0.110*		+0.002	−0.198*
direct vs. pivot		+0.004	−0.109*		+0.081*	−0.266*

Table 10: Paired differences from the released text on 600 items (95% bootstrap intervals within language). "Places > neutral" is the share of items on which the places-only variant is scored strictly more specific than the neutral one.

	Qwen3-235B	**Gemma-4-31B**
Specificity: neutral − released	−2.58	−2.27
Specificity: names − released	−2.32	−2.05
Specificity: places − released	−0.14 [−0.24, −0.04]	−0.44
Specificity: noplaces − released	−1.25	−1.07
Specificity: explained − released	≈ 0	≈ 0
Regional fit: mismatched − released	−1.59	−2.31
P(places > neutral) / P(names > neutral)	0.97 / 0.30	0.89 / 0.23
Released judged "substantive"	0.41	0.58

C.4 Are the bank records right?

Qwen3-235B with reasoning and Gemma-4-31B independently audit all 673 unit, scenario and rate records, checking every conversion and flagging figures they believe wrong for the stated place and time. Alone, each supports a different story: Gemma finds 51% of records fine and 11% majorly wrong, Qwen 25% and 44%, and they agree on severity for 33% of records ($\kappa = 0.11$), the disagreement concentrated in time-varying prices. What both flag is small: 29 records (4.4%) are major under both judges. Reading the 38 records that are major under both or share a disputed conversion, three kinds recur: the tola listed at the 10 g bullion convention beside a traditional 11.66 g bhori; magnitude errors such as a sugarcane yield of 80–100 tonnes per acre; and state wage rates one revision out of date. The first two are corrected in the next version and the third needs a year attached. Agreement between two models prioritises review; it does not certify the other records. An audit against primary sources remains owed.

From records to items. Two kinds of traceability should be kept apart. *Source and formula lineage* (source problem, formula, operands, every intermediate text) is complete for every item. *Context-record provenance* is what the bank's correctability rests on, and released items do not carry it: they store no record identifier. It is recoverable from the lineage, because every job stores verbatim the context block its prompt carried. Matching those blocks by content to the 673 audited records, all 88,066 released items match at least one record; all come from one Stage 1 run and hence one bank version; and we release, per item, the matched records and a hash of the block. The result separates what was *supplied* from what was *used*. Supplied: a prompt carries a units block, a scenario record, several scenario options and a universal-rate row, and the broadcast rows are widely shared, so 77% of items (6,846 of 8,860 test items) had at least one of the 29 doubly-flagged records somewhere in their prompt. Used: the generator names its scenario on the first line of its output; that name is one the prompt offered for 51% of items (otherwise the generator wrote its own) and matches an audited

record for 45%. By that measure 1,797 items (2.0%; 168 test items) are set in a doubly-flagged scenario and 2,488 (2.8%; 248 test) in a scenario with any shared flag. Among audited test items those set in a flagged scenario are no less often valid (0.85, $n = 37$, against 0.82), as expected when operands come from the contract and not from the record, but validity does not measure whether a stated price or conversion is right. For the items whose scenario line matches no record we know what was supplied and not what was used: the lineage establishes that the supplied conditioning is explicit and recoverable, not that every item uses it. Correcting the workbook repairs none of these items; the listed ones are to be re-checked or regenerated, and the next version writes record identifiers and the chosen option into each item.

D Item validity

A gold-blind census by two model families. The planned audit uses native-speaker annotators; we ran the same protocol with model reviewers, so it measures what two independent model families agree on. Over the whole public test split (8,860 items), Gemma-4-31B and Qwen3-235B with reasoning each write, gold-blind, the expression they believe the question asks for, judge whether it is well posed, and list defects, three passes each; a label is assigned in code. A majority verdict is returned by Gemma for 8,794 items, by Qwen for 8,846 and by both for 8,781; the other 79 produced no parseable verdict from one family and ship unlabelled. The families agree on 84.1% of the 8,781; the 1,395 disagreements go to a third gold-blind adjudication that sees both reviews in random order, and three stay undecided, which leaves 8,778 labelled items. All shares below are of labelled items; counting every unlabelled item as not valid would put the valid share at 81.9% of the split. 82.7% of items are well posed with a gold an independent reviewer re-derives; ambiguity or underspecification is the dominant defect (13.2%): a count that comes out non-integral, an unstated unit, a non-standard conversion. 1.9% have a well-posed question whose answer the census believes differs from the stored gold and 2.2% alter an operand, so 4.1% of items have a gold that a careful reader would not reach. A single family brackets the adjudicated figure (88.8% valid for Gemma alone, 80.2% for Qwen alone). Validity is flat across languages (81.5–83.8%) and falls with complexity, from 88.3% at the lowest level to 74.6% at the highest ($n = 260$). Table 11 gives the label shares and, for all seven models, accuracy on the full split and on the valid core: every model gains and the order is the same, with the same adjacent pairs of the top five unseparated in both. The census labels ship with every labelled item. The valid core is model-adjudicated: it has not been calibrated against native-speaker judgements.

Table 11: Validity census of the *whole* public test split (8,778 of 8,860 items labelled; 79 lack a majority verdict from one family and 3 stay undecided). Left: label shares from each reviewer family alone (majority of three gold-blind passes) and after blind adjudication of the 1,395 cross-family disagreements, of which 1,392 are resolved (bold); the families agree on 84.1% of items, and the two single-family valid shares bracket the adjudicated one. Validity is flat across languages (0.815–0.838) and falls with complexity (0.882 at level 1 to 0.746 at level 4, $n = 260$). Right: zero-shot accuracy on all 8,860 test items and on the 7,258 census-valid items. The ordering is identical position for position, and the same adjacent pairs are separated by a language-clustered bootstrap in both columns: the three tiers (five leading models, Gnani Evon 3.3, Nemotron) do not depend on the flagged items. Labels ship per item.

Label	Gemma-4-31B	Qwen3-235B	adjudicated
valid	0.888	0.802	**0.827**
ambiguous / underspecified	0.056	0.148	**0.132**
operand altered	0.026	0.028	**0.022**
answer differs	0.029	0.022	**0.019**
unit inconsistency	0.001	0.001	**0.001**
language defect	0.000	0.000	**0.000**

Model	all items	valid core	shift
Gemma-4-31B	0.871	0.910	+0.039
Sarvam-105B	0.865	0.908	+0.043
Gemma-4-26B	0.843	0.884	+0.042
Qwen3.8-27B	0.828	0.879	+0.051
Sarvam-30B	0.823	0.872	+0.049
Gnani Evon 3.3	0.661	0.705	+0.044
Nemotron-3-Nano	0.262	0.283	+0.022

An earlier stratified sample of 1,177 items under the same protocol gave 81.6% valid and language estimates from 0.75 to 0.88; the census shows that spread was sampling noise at about 100 items per language. The sample's 81.6% and the census's 82.7% are two assessments of the same release: no item was repaired between them. On that sample the families agree on validity at $\kappa = 0.42$, which is why a third pass adjudicates and why a native-speaker calibration of the labels remains owed. Table 13 repeats the interventions of §6 on audit-valid items of the sample; with the census the same conditioning runs on the whole 3,000-item subsample: pooled effects move by at most 1.2 points and

Table 12: Intervention effects on all items and on the census-valid core (pooled over seven models, reasoning off, paired against zero-shot on the same items unless stated; language-clustered bootstrap, 2,000 resamples; * interval excludes zero). *Shift* is the valid-core effect minus the all-items effect on shared resamples. No pooled effect changes sign or loses significance on the valid core, and the largest shift is 1.2 points.

Intervention	n	n valid	all items	valid core	shift
distractor sentence	2,747	2,239	-0.070 [-0.080, -0.060]*	-0.072 [-0.084, -0.061]*	-0.002 [-0.006, +0.001]
romanised question	2,962	2,417	-0.134 [-0.163, -0.105]*	-0.140 [-0.170, -0.110]*	-0.006 [-0.008, -0.004]*
code-mixed question	1,786	1,439	+0.040 [+0.028, +0.056]*	+0.044 [+0.028, +0.064]*	+0.004 [-0.000, +0.009]
paraphrase	2,698	2,197	-0.004 [-0.012, +0.002]	-0.001 [-0.008, +0.006]	+0.004 [+0.001, +0.006]*
oracle English input	2,943	2,403	+0.113 [+0.086, +0.141]*	+0.124 [+0.094, +0.155]*	+0.011 [+0.006, +0.015]*
English input, program register	2,943	2,403	+0.104 [+0.077, +0.129]*	+0.115 [+0.085, +0.144]*	+0.012 [+0.007, +0.017]*
program we execute	2,943	2,403	+0.005 [-0.008, +0.018]	+0.009 [-0.005, +0.022]	+0.003 [+0.000, +0.006]*
calculator	2,943	2,403	-0.005 [-0.020, +0.009]	-0.003 [-0.018, +0.013]	+0.003 [-0.001, +0.006]

no model's by more than 3.1 (Table 12). On this model-adjudicated valid core the distractor costs 7.2 points pooled, romanised input 14.0 (reasoning off) and 14.6 (on), code-mixing *helps* by 4.4 and 3.8, and paraphrase is -0.1 (n.s.).

A single gold-blind model reader gives a third view: shown only the question text of all 8,860 test items, it reconstructs an expression equal to the hidden gold on 79.1%, flat across languages; where it disagrees the seven solvers mostly box the gold, siding with the reader on 50 items (0.56%).

Table 13: Do the gaps survive on audit-valid items? Each cell is the paired change, in points, on the full evaluation set / on all audited items scored in that arm / on the audit-valid items among those (the last two weighted to the public test; 900–1,177 audited and 733–960 valid items, or 368–399 and 298–322 for arms marked †, which ran on the 3,000-item subset). The first step changes sample membership and the second changes validity alone. *: a 95% interval on valid minus audited, resampling audited items, excludes zero; °: the valid-only interval contains zero. Across all 50 arm-by-model cells (seven models) filtering moves an effect by at most 3.0 points, and in the 5 cells where the move is distinguishable from zero the effect grows. Large English-input and numeral-script effects persist; effects under about three points and several † arms are not resolved at this sample size, and where an estimate moves between the first and last number (Qwen3.8, distractor and program of thought) it is the audited sample, not validity, that moves it. The last row is the best-minus-worst language spread on all audited / audit-valid items (about 80 items per language, so both are noisier than the full-split spreads of Table 3); no spread shrinks when flagged items are removed. The translated-Bengali comparison cannot be repeated this way: only some thirty audited items are Bengali items in that arm.

Intervention	Gemma-4-31B	Sarvam-30B	Qwen3.8-27B	Gnani Evon 3.3	Nemotron-3-Nano
Indian-English intermediate	+5.1 / +3.5 / +3.0	+9.6 / +8.1 / +8.8	+9.6 / +9.9 / +8.7	+19.2 / +18.7 / +19.4	+52.1 / +51.6 / +53.3*
oracle English translation†	+1.6 / +1.5 / +3.7	+6.7 / +8.9 / +9.1	+6.3 / +7.4 / +8.7	+17.1 / +18.5 / +19.8	+46.8 / +46.8 / +49.8*
digits to ASCII	+1.3 / +0.8 / +0.3°	+3.8 / +2.6 / +3.2	+1.4 / +1.4 / +1.9	+4.3 / +4.0 / +4.5	+5.4 / +6.0 / +5.9
digits to native script	−4.9 / −3.9 / −4.2	−14.0 / −13.3 / −13.7	−0.9 / −0.8 / −1.3°	−8.1 / −6.7 / −6.4	−7.6 / −8.5 / −9.1
native-numeral answer	−7.5 / −7.3 / −8.1	−20.7 / −22.6 / −23.6	−2.5 / −2.4 / −3.4	−3.0 / −3.6 / −3.4	−3.4 / −1.4 / −3.2*
one distractor sentence†	−4.7 / −2.7 / −4.0	−8.6 / −5.5 / −6.4	−6.6 / −2.3 / −1.2°	−9.2 / −7.0 / −7.2	−6.2 / −6.5 / −8.0
romanised input†	−3.7 / −3.0 / −2.4°	−12.5 / −11.0 / −11.8	−21.2 / −16.4 / −16.4	−20.0 / −13.2 / −13.0	−16.4 / −17.1 / −19.3*
program of thought†	+4.8 / +4.5 / +4.9	+0.5 / −0.4 / −1.2°	−1.7 / +3.2 / +4.5°	−5.9 / −7.6 / −7.8	+1.2 / +2.5 / +0.7°
language spread, zero-shot	0.204 / 0.223	0.256 / 0.299	0.173 / 0.189	0.660 / 0.658	0.628 / 0.679

Where items break. Every released item passes through three texts: the English problem written by Call A, its literal translation, and the re-voiced question we ship. We give each, as a separate request, to the blind reader (Qwen3-235B, reasoning off); an item is *valid* at a stage when the reader's expression, evaluated in code, equals the stored gold. Translation costs 6.9 points of validity and re-voicing recovers 2.0, or 29% of the loss (Table 14). Every language loses in translation, from 3.9 points (Hindi) to 12.9 (Assamese), and re-voicing wins back most where translation hurts most: +8.3 in Assamese, +4.9 in Odia, +3.5 in Tamil. A second reader family (Gemma-4-31B) is more lenient in absolute terms and reproduces every stage contrast in sign and significance. "Invalid" here means "this reader does not reconstruct the gold": 16.0% of *English* drafts are already not reconstructed, which bounds reader error plus Call A defects from above.

Do the gates catch broken items? We take 1,020 released items and break them in known ways: one operand changed, one deleted, two swapped where that changes the answer, and two semantic mutations written by Gemma-4-31B (a relation flipped; a unit changed), kept only when every number is untouched and the stated new answer differs from gold. Untouched controls measure how often

Table 14: Validity of the same 8,725 public-test items at three pipeline stages, read blind by one reader. "Operands" means the reader's expression uses exactly the contract's formula operands. Stage differences are paired, with 95% intervals from 10,000 bootstrap resamples of items within language. The released-stage read is a within-run replicate of the reader study of §D. The 135 test items (1.5%) whose intermediate draft states the answer are excluded from all three stages; no such statement survives into a released question.

	Valid			Operands match		
Items	Call A (en)	translation	released	Call A (en)	translation	released
All (8,725)	0.840	0.771	0.791	0.747	0.726	0.735
difference from previous stage		-0.069^*	$+0.020^*$		-0.021^*	$+0.009^*$
interval		[−0.077, −0.061]	[+0.012, +0.028]		[−0.029, −0.014]	[+0.002, +0.016]
Tier 1 (1,209)	0.928	0.881	0.910			
Tier 2 (5,626)	0.845	0.772	0.794			
Tier 3 (1,890)	0.768	0.698	0.706			
Repaired after Call B (578)	0.785	0.706	0.772			

a gate objects to an unedited item (Table 15). The shipped numeral check catches every changed or deleted operand, never objects to a control, and is blind to a swap and to both semantic mutations, where every number is still on the page. Those are caught by a model that re-derives the answer: a blind reader or a solver notices 93% of swaps, 87% of unit changes and 80% of flipped relations. The price is disagreement on untouched items: there the reader fails to recover gold 27.2% of the time. That figure mixes reader error with defects the items already had (the census above finds 17% of items not fully valid), so it is not a false-alarm rate; a reader used as a hard filter would discard up to a quarter of items, and requiring both models to object lowers the rate to 5.8%. The released build applied the numeral check only; the model readers were run offline as diagnostics, and the next release exports an item only when neither reader objects.

Table 15: Share of items flagged by each gate: G1, the shipped numeral-presence check; G2, a gold-blind Qwen3-235B reader; G3, a Gemma-4-31B zero-shot solver disagreeing with gold. The control row is the rate of objections to untouched released items; every other row is a detection rate on items whose stored gold is no longer the answer. For the two semantic rows the change of answer is the mutator's own claim, so those rates are lower bounds.

mutation	n	G1	G2	G3	G1 ∨ G2	G2 ∨ G3	G2 ∧ G3
Control (untouched)	1,020	.000	.272	.078	.272	.292	.058
One operand changed	1,009	1.000	.985	.989	1.000	.993	.981
One operand deleted	1,020	1.000	.944	.905	1.000	.951	.898
Two operands swapped	669	.000	.910	.873	.910	.931	.852
Relation flipped (semantic)	794	.000	.710	.674	.710	.801	.583
Unit changed (semantic)	833	.000	.816	.688	.816	.866	.639

Are the stored formulas right? For each of 6,155 distinct source problems of the public test split, Qwen3-235B writes its own expression from the problem text and variable roles alone. Both expressions are evaluated by a whitelisted-AST interpreter at the source valuation and at 20 random valuations. For 88.0% of the 5,764 problems with a usable reference the two are functionally identical (Table 16). Where they differ at the source, the source dataset's answer sides with the stored formula four times in five; where they agree at the source and differ elsewhere, three quarters are a modelling difference in which the stored formula is the more general. This audit flags 158 of the 5,764 sources with a usable reference (2.7%). That is a flagged share, not a bound: two model-derived formulas can share an error, and the 391 sources (6.4%) without a usable reference are not covered: for 222 the reference generation reached its token cap, for 134 no expression could be parsed from it, and for 35 the expression did not evaluate. They are 7.4% of AQuA-RAT sources, 5.4% of MathQA, 2.0% of GSM8K and none of SVAMP, so the uncovered sources lean towards the harder source sets. Zero-shot accuracy on the 228 test items descended from flagged blueprints is within five points of the rest for every model, which does not by itself show that the flag over-counts. The identifiers ship with the release.

Table 16: Stored blueprint formula against an independent re-derivation by Qwen3-235B (5,764 problems with a usable reference).

outcome	problems	share
equivalent at the source and at 20 random valuations	5,071	0.880
agree at the source, differ elsewhere	389	0.067
reference reads a variable the stored formula never reads	291	0.050
same variables	98	0.017
differ at the source valuation	304	0.053
stored formula reproduces the source dataset's answer, reference does not	243	0.042
reference reproduces it, stored formula does not	41	0.007
neither (19) or both within tolerance (1)	20	0.003

E Evaluation protocol and extended results

Protocol. Seven instruction-tuned models are evaluated, plus two further rungs of the Gemma-4 family for the scale ladder. All generations are produced offline with vLLM in bf16, maximum model length 40,960, with reasoning switched through the chat template and one fixed seed for every sampled arm, subsample and shuffle. The zero-shot prompt is the question, one sentence naming its language and asking for a solution, and five format rules, the last requiring the final answer in `\boxed{}`. Reasoning-off arms are greedy with a 4,096-token cap; reasoning-on arms have a 30,000-token cap and are run greedy, at $T{=}0.6$, and with the vendor's recommended sampling. Both Sarvam checkpoints are degenerate under greedy decoding, so every Sarvam number is produced with sampled decoding ($T{=}1.0$, top-p 0.95, one sample) and a Sarvam paired difference carries sampling noise on both sides. A response is correct iff the value in its last box is within 10^{-4} relative error of the gold or equals it after both are rounded to two decimals, the precision at which golds are stored; native-script digits are normalised before parsing. We keep three things apart. A *cap-hit* is a generation that reaches its token cap; *truncated* is a cap-hit with no parseable box, which is scored wrong; its share is tabulated for the reasoning-on arms of Tables 36, 33 and 34 and is in the released per-item records for every other arm; a cap-hit that still contains a box is scored on it, and a completed response with no box is wrong but not truncated. Table 34 splits by cap-hit. Full-split arms use the 8,860 public-test items; the more expensive arms use a fixed 3,000-item subsample, 250 per language. Uncertainty is a language-stratified item bootstrap with 10,000 draws on paired differences; a starred difference is one whose 95% interval excludes zero. Whenever a model reads, judges or rewrites an item it is Qwen3-235B, Gemma-4-31B or DeepSeek-V4-Flash, never the model under test in that role. Model-written Python is never executed; programs are interpreted by a restricted evaluator over a whitelist of arithmetic nodes. The column "Evon 3.3" is Gnani Evon 3.3.

E.1 Headline numbers, decoding and translated benchmarks

Table 17: Headline zero-shot accuracy on the public test split under different decoding regimes (paired difference against the thinking-off greedy row). Sarvam columns are sampled throughout (see text): their "greedy" cells are one sample at $T{=}1.0$, top-p 0.95, and with thinking on the token cap rises from 16,384 to 30,000. Their one-point gain from thinking is of the size of the fall in truncation (Sarvam-105B: 2.3% $\rightarrow$ 1.2%). At $T{=}0.6$ Sarvam-30B truncates 16.1% of generations (mean 13,222 tokens against 3,937), the same loop that makes it degenerate under greedy decoding; Sarvam-105B shows a milder form of it (4.4% against 1.1% under the vendor setting).

arm	G-26B-A4B	G-31B	Nemotron	Qwen3.8	Evon 3.3	S-105B	S-30B
thinking off, greedy	0.843	0.871	0.262	0.828	0.661	0.865	0.823
thinking on, greedy	0.624 (−0.219)†	0.817 (−0.053)†	0.589 (+0.328)†	0.922 (+0.094)†	0.828 (+0.166)†	0.875 (+0.010)†	0.832 (+0.008)†
thinking on, $T{=}0.6$, $n{=}5$	0.805	0.902	0.591	0.926	0.846	0.859	0.734
thinking on, vendor, $n{=}5$	0.881	0.921	0.591	—	0.843	0.873	0.820
answer only (no working)	0.618 (−0.225)†	0.896 (+0.025)†	0.106 (−0.156)†	0.342 (−0.486)†	0.232 (−0.429)†	0.850 (−0.015)†	0.820 (−0.004)

Table 20 is per-item (micro) accuracy within each split, while the full-corpus column of Table 3 is the mean of the twelve per-language accuracies (macro). The two agree within 0.002 for every model except Nemotron, whose weakest languages are slightly under-represented: 0.6110 micro (the item-weighted mean of 61.08, 60.52 and 61.88) against 0.6033 macro.

Table 18: Head-to-head with translated benchmarks, reasoning off, same prompt, grader and decoding. BharatMath columns are 250 items per language. English is saturated; BharatMath-Bengali re-orders the models (Spearman ρ between MGSM-bn and BharatMath-bn $= 0.46$).

	Gemma 31B	**Gemma 26B**	**Sarvam 105B**	**Sarvam 30B**	**Qwen3.8 27B**	**Gnani Evon 3.3**	**Nemotron 30B**
MGSM English	0.984	0.960	0.988	0.964	0.976	0.952	0.936
MGSM Bengali	**0.916**	0.888	0.880	0.844	0.828	0.792	0.472
MGSM Telugu	**0.896**	0.876	0.852	0.848	0.656	0.828	0.340
MSVAMP Bengali	**0.813**	0.798	0.807	0.770	0.799	0.786	0.608
BharatMath Bengali	0.784	0.752	**0.864**	0.852	0.788	0.724	0.276
BharatMath Telugu	**0.888**	0.872	0.810	0.696	0.752	0.704	0.276

Table 19: Bengali with the source mix held fixed. MGSM is translated GSM8K; BharatMath draws on five source sets. Restricting BharatMath-Bengali to its 62 test items descended from GSM8K (last column, 95% item bootstrap) keeps the two Sarvam models first and the ordering unlike MGSM-Bengali's (Spearman $\rho = 0.21$; 0.32 for all 788 Bengali test items, which are the full split and not the 250-item subsample of Figure 5). With 62 items most intervals overlap, and the items still differ from MGSM in operands, selection and filtering history, so this removes one alternative explanation and not the others.

Model	**MGSM-en**	**MGSM-bn**	**BharatMath-bn, all**	**BharatMath-bn, GSM8K-derived**
Gemma-4-31B	0.984	0.916	0.777	0.806 [0.71, 0.90]
Gemma-4-26B	0.960	0.888	0.744	0.742 [0.63, 0.85]
Sarvam-105B	0.988	0.880	0.876	0.903 [0.82, 0.97]
Sarvam-30B	0.964	0.844	0.858	0.919 [0.85, 0.98]
Qwen3.8-27B	0.976	0.828	0.843	0.790 [0.69, 0.89]
Gnani Evon 3.3	0.952	0.792	0.763	0.855 [0.77, 0.94]
Nemotron-3-Nano	0.936	0.472	0.299	0.468 [0.34, 0.60]

Table 20: Accuracy (%) by split, original full-corpus predictions; difference in points with a 95% interval.

model	train	public test	private test	public − private [95% CI]
Gemma-4-31B	87.09	87.15	87.12	+0.03 [−1.08, +1.05]
Gemma-4-26B-A4B	85.15	84.96	85.29	−0.34 [−1.44, +0.91]
Qwen3.8-27B	94.07	93.48	94.18	−0.70 [−1.48, +0.02]
Sarvam-105B	86.01	85.57	86.21	−0.64 [−1.74, +0.36]
Sarvam-30B	81.96	81.29	82.29	−1.01 [−2.16, +0.09]
Nemotron-3-Nano	61.08	60.52	61.88	−1.36 [−2.90, +0.24]
Evon 3.3	84.83	84.54	84.93	−0.39 [−1.57, +0.68]

Table 21: Stability of rankings under language-stratified subsampling (300 draws per size).

items / language	total	mean τ (models)	5th pct. τ	draws with an inversion	mean τ (languages)
25	300	0.82	0.62	90%	0.49
100	1,200	0.92	0.81	60%	0.63
250	3,000	0.96	0.90	35%	0.72
500	6,000	0.99	0.90	8%	0.84

Table 22: Language effect on accuracy relative to Hindi with a fixed effect per source problem: the shift common to all models, its standard deviation across models, the raw gap, and each model's remainder.

	common	sd	raw gap	G-31B	G-26B-A4B	Qwen3.8	S-105B	Evon 3.3	S-30B	Nemotron
as	−0.074	0.108	−0.098	−0.144	−0.124	−0.010	+0.033	−0.011	+0.003	−0.265
bn	−0.054	0.077	−0.057	−0.141	−0.161	−0.034	+0.031	+0.039	−0.076	−0.035
gu	−0.042	0.131	−0.067	−0.050	−0.027	−0.009	+0.044	+0.058	+0.016	−0.326
kn	−0.048	0.102	−0.048	−0.012	−0.012	−0.006	−0.009	+0.008	−0.025	−0.278
ml	−0.160	0.170	−0.164	−0.054	−0.073	−0.048	−0.063	−0.075	−0.348	−0.462
mr	−0.040	0.090	−0.054	−0.009	−0.052	−0.019	+0.057	−0.017	−0.008	−0.230
or	−0.146	0.338	−0.163	+0.010	−0.050	+0.004	+0.026	−0.117	+0.010	−0.904
pa	−0.018	0.095	−0.051	−0.017	+0.003	−0.037	+0.073	+0.009	+0.055	−0.214
ta	−0.041	0.138	−0.060	+0.016	−0.025	−0.006	+0.013	+0.073	−0.012	−0.347
te	−0.069	0.101	−0.101	−0.008	+0.007	−0.030	−0.054	−0.017	−0.094	−0.286
ur	+0.003	0.029	−0.012	+0.013	−0.026	−0.020	+0.061	−0.002	+0.000	−0.004

E.2 Gap 1: language of the question

Table 23: Language decomposition, thinking off (paired difference against the same model's zero-shot on the same items). B1, B2, B6 and the native-numeral answer arm use the full test split; B3 and B4 use the 3,000-item subsample, whose baseline is the second row. B2 shows the model the intermediate English draft; dropping the items whose draft states the answer (Table 14) moves no model's B2 effect by more than 0.4 points.

arm	G-26B-A4B	G-31B	Nemotron	Qwen3.8	Evon 3.3	S-105B	S-30B
zero-shot, full split	0.843	0.871	0.262	0.828	0.661	0.865	0.823
zero-shot, subsample	0.840	0.863	0.265	0.818	0.646	0.849	0.809
B1 English instruction	0.845 (+0.003)	0.897 (+0.027)†	0.347 (+0.086)†	0.852 (+0.024)†	0.763 (+0.101)†	0.864 (−0.001)	0.821 (−0.003)
B2 Indian-English intermediate	0.848 (+0.005)	0.922 (+0.051)†	0.788 (+0.521)†	0.926 (+0.096)†	0.859 (+0.192)†	0.914 (+0.049)†	0.922 (+0.096)†
B3 self-translate, solve	0.822 (−0.018)†	0.885 (+0.022)†	0.439 (+0.174)†	0.844 (+0.026)†	0.794 (+0.148)†	0.833 (−0.017)†	0.736 (−0.073)†
B4 oracle English	0.819 (−0.020)†	0.878 (+0.016)†	0.733 (+0.468)†	0.882 (+0.063)†	0.818 (+0.171)†	0.875 (+0.025)†	0.877 (+0.067)†
B6 first step given	0.887 (+0.044)†	0.904 (+0.033)†	0.344 (+0.083)†	0.861 (+0.033)†	0.728 (+0.067)†	0.885 (+0.020)†	0.864 (+0.041)†
answer in native numerals	0.788 (−0.055)†	0.797 (−0.074)†	0.227 (−0.034)†	0.803 (−0.025)†	0.632 (−0.030)†	0.765 (−0.099)†	0.616 (−0.207)†

Table 24: Matched language × culture design: accuracy on 600 contracts rendered four ways with identical mathematics and gold, and paired contrasts with 95% bootstrap intervals (4,000 resamples of contracts within language).

	Accuracy				Regional − source		Indic − English	
Model	E_{src}	E_{reg}	L_{src}	L_{reg}	in English	in Indic	source	regional
Gemma-4-31B	.848	.847	.703	.790	-0.002 [-0.028, $+0.027$]	$+0.087$ [$+0.048$, $+0.125$]	-0.145 [-0.178, -0.113]	-0.057 [-0.082, -0.033]
Sarvam-105B	.772	.735	.642	.695	-0.037 [-0.072, -0.002]	$+0.053$ [$+0.013$, $+0.095$]	-0.130 [-0.162, -0.097]	-0.040 [-0.068, -0.012]
Gemma-4-26B-A4B	.792	.782	.695	.790	-0.010 [-0.042, $+0.022$]	$+0.095$ [$+0.057$, $+0.135$]	-0.097 [-0.133, -0.062]	$+0.008$ [-0.020, $+0.037$]
Qwen3.8-27B	.867	.842	.652	.773	-0.025 [-0.053, $+0.005$]	$+0.122$ [$+0.082$, $+0.163$]	-0.215 [-0.250, -0.180]	-0.068 [-0.098, -0.040]
Sarvam-30B	.733	.715	.607	.687	-0.018 [-0.053, $+0.018$]	$+0.080$ [$+0.037$, $+0.125$]	-0.127 [-0.162, -0.092]	-0.028 [-0.062, $+0.005$]
Gnani Evon 3.3	.798	.808	.543	.640	$+0.010$ [-0.022, $+0.043$]	$+0.097$ [$+0.053$, $+0.140$]	-0.255 [-0.293, -0.217]	-0.168 [-0.205, -0.133]
Nemotron-3-Nano	.725	.710	.222	.282	-0.015 [-0.050, $+0.020$]	$+0.060$ [$+0.020$, $+0.102$]	-0.503 [-0.545, -0.463]	-0.428 [-0.470, -0.387]

The matched comparison under one gate, and MGSM regenerated. Table 24 compares a gated regional cell with an unfiltered translated one. Table 25 removes that asymmetry: a contract enters only if every operand survives as a digit in *both* Indic cells, which keeps 527 of 600 (the translator loses a number in 12.2% of contracts, from 2% in Marathi to 26% in Punjabi). The right block takes the 203 MGSM problems that pass our blueprint filter with the holdout gate lifted for this experiment only. Stage 1 routes each problem to the language cells compatible with its domain (3.6 languages per problem on average), which gives 723 problem–language pairs with both renderings scored, and the block compares a literal translation with our regenerated rendering of the *same* operands and gold in the same language. The production filter drops every GSM8K-test source before generation, so no released item descends from an MGSM problem.

Table 25: Two source-matched comparisons. Left, the language-by-framing 2 × 2 of Table 24 restricted to the 527 of 600 contracts on which the machine translation of the source keeps every number (IndicTrans2 loses one on 12.2% of problems, from 2% to 26% by language), so that both Indic cells pass the same numeral check. Regional framing costs at most five points in English (two models, the rest n.s.) and in the Indic languages the regional item is solved *more* often than the translated source (pooled, and for five of seven models; never significantly less); the interaction is positive for every model. Right, 723 problem–language pairs from 203 MGSM source problems with the *source* operands kept, rendered by literal translation and by the full pipeline: regeneration makes the same mathematics harder for every model. The census does not cover these renderings, so the loss is not apportioned between richer text and rendering defects; the comparison shows that BharatMath items are not an easier restatement of their sources. MGSM cannot enter the corpus: it is drawn from the GSM8K test split, which a Stage 0 gate excludes as a source. * interval excludes zero.

	language × framing, number-preserving contracts							source operands, two renderings		
Model	En., source	En., regional	Indic, translated	Indic, regional	regional − source, En.	regional − translated, Indic	interaction	literal	pipeline	pipeline − literal
Gemma-4-31B	0.858	0.852	0.774	0.801	-0.006 [-0.034, +0.028]	+0.027 [+0.007, +0.044]*	+0.032 [-0.009, +0.067]	0.965	0.807	-0.159 [-0.190, -0.131]*
Sarvam-105B	0.793	0.744	0.702	0.698	-0.049 [-0.089, -0.011]*	-0.004 [-0.048, +0.042]	+0.045 [+0.002, +0.094]*	0.931	0.786	-0.144 [-0.177, -0.109]*
Gemma-4-26B	0.810	0.789	0.765	0.803	-0.021 [-0.057, +0.017]	+0.038 [+0.002, +0.073]*	+0.059 [+0.009, +0.109]*	0.945	0.799	-0.144 [-0.173, -0.117]*
Qwen3.8-27B	0.877	0.844	0.719	0.784	-0.032 [-0.056, -0.007]*	+0.065 [+0.024, +0.100]*	+0.097 [+0.056, +0.136]*	0.858	0.749	-0.106 [-0.140, -0.072]*
Sarvam-30B	0.757	0.729	0.664	0.698	-0.029 [-0.074, +0.008]	+0.034 [-0.015, +0.082]	+0.063 [+0.019, +0.105]*	0.922	0.769	-0.153 [-0.198, -0.111]*
Gnani Evon 3.3	0.799	0.810	0.588	0.649	+0.011 [-0.017, +0.045]	+0.061 [+0.024, +0.096]*	+0.049 [-0.002, +0.099]	0.805	0.689	-0.118 [-0.159, -0.080]*
Nemotron-3-Nano	0.732	0.719	0.245	0.292	-0.013 [-0.053, +0.027]	+0.047 [+0.017, +0.079]*	+0.061 [+0.007, +0.120]*	0.420	0.350	-0.069 [-0.108, -0.034]*
pooled, seven models	0.804	0.784	0.637	0.675	-0.020 [-0.045, +0.006]	+0.038 [+0.018, +0.058]*	+0.058 [+0.033, +0.083]*	0.835	0.707	-0.128 [-0.151, -0.105]*

E.3 Gap 2: numeral script

Table 26: Digit-script intervention, thinking off, full test split (paired difference against the released questions).

arm	G-26B-A4B	G-31B	Nemotron	Qwen3.8	Evon 3.3	S-105B	S-30B
all digits ASCII	0.862 (+0.019)†	0.883 (+0.013)†	0.315 (+0.054)†	0.842 (+0.014)†	0.704 (+0.043)†	0.880 (+0.015)†	0.861 (+0.038)†
all digits native	0.813 (−0.029)†	0.822 (−0.048)†	0.186 (−0.076)†	0.819 (−0.009)†	0.581 (−0.081)†	0.784 (−0.080)†	0.684 (−0.140)†

Table 27: Digit-swap twins. Each twin is byte-identical to its original except for the digit script; 3,295 native-digit renderings are rewritten to ASCII and 2,705 ASCII-digit renderings to the language's own digits. Δ is twin − original with a 95% bootstrap interval over contracts. All sixteen intervals exclude zero and all sixteen signs are as predicted.

	native → ASCII ($n = 3{,}295$)			ASCII → native ($n = 2{,}705$)		
model, mode	original	twin	Δ	original	twin	Δ
Gemma-4-31B, off	0.792	0.812	$+0.020$ [+0.006, +0.034]	0.816	0.704	-0.112 [-0.147, -0.079]
Gemma-4-26B-A4B, off	0.747	0.799	$+0.053$ [+0.035, +0.072]	0.793	0.715	-0.078 [-0.105, -0.052]
Qwen3.8-27B, off	0.773	0.806	$+0.033$ [+0.018, +0.050]	0.746	0.724	-0.023 [-0.041, -0.005]
Sarvam-30B, off	0.738	0.814	$+0.076$ [+0.053, +0.100]	0.811	0.502	-0.309 [-0.355, -0.264]
Evon 3.3, off	0.541	0.632	$+0.090$ [+0.067, +0.115]	0.664	0.512	-0.152 [-0.184, -0.122]
Evon 3.3, on	0.752	0.790	$+0.038$ [+0.022, +0.054]	0.810	0.727	-0.082 [-0.109, -0.057]
Nemotron-3-Nano, off	0.162	0.251	$+0.089$ [+0.069, +0.111]	0.272	0.112	-0.160 [-0.190, -0.131]
Nemotron-3-Nano, on	0.441	0.596	$+0.155$ [+0.125, +0.186]	0.642	0.300	-0.342 [-0.384, -0.301]

E.4 Gap 3: execution, and scale

Table 28: Program of thought (PoT) against natural-language solving (NL), thinking off, 3,000 items, paired. Spread is the best-minus-worst language accuracy. Sarvam rows are sampled.

model	executable	PoT	NL	paired Δ	both wrong	PoT only	NL only	spread NL → PoT
Gemma-4-31B	0.978	0.911	0.863	+0.048	0.054	0.083	0.035	0.179 → 0.072
Gemma-4-26B-A4B	0.966	0.866	0.840	+0.026	0.070	0.090	0.064	0.167 → 0.080
Sarvam-105B	0.976	0.867	0.849	+0.018	0.076	0.075	0.057	0.125 → 0.180
Sarvam-30B	0.965	0.814	0.809	+0.005	0.114	0.077	0.072	0.312 → 0.360
Qwen3.8-27B	0.968	0.801	0.818	−0.017	0.108	0.074	0.091	0.217 → 0.241
Evon 3.3	0.887	0.587	0.646	−0.059	0.220	0.134	0.192	0.528 → 0.380
Nemotron-3-Nano	0.888	0.277	0.265	+0.012	0.625	0.110	0.098	0.608 → 0.588

Table 29: What happens to natural-language failures when the model writes a program instead (3,000 items, reasoning off). For each item the model answered wrongly in natural language, its program is either *right* (executes to the gold: the formalisation was available and the natural-language computation failed), *wrong value* (executes to something else: the formalisation itself is wrong), or *not executable*. The interpreter does the arithmetic, so an evaluation error inside a correct expression cannot be observed, and the program prompt may itself change how the question is read; the split is evidence about where failures sit, not a causal decomposition. Last column: share of natural-language successes whose program is wrong.

					program, among NL failures			**NL right,**
Model	n	**NL acc.**	**program acc.**	**NL failures**	right	wrong value	not exec.	**program wrong**
Gemma-4-31B	3000	0.863	0.911	411	0.61	0.34	0.05	0.041
Gemma-4-26B	3000	0.840	0.866	480	0.56	0.37	0.07	0.076
Sarvam-105B	3000	0.849	0.867	452	0.50	0.44	0.06	0.068
Sarvam-30B	3000	0.809	0.814	573	0.40	0.50	0.10	0.089
Qwen3.8-27B	3000	0.818	0.801	546	0.41	0.51	0.08	0.112
Gnani Evon 3.3	3000	0.646	0.587	1062	0.38	0.51	0.12	0.298
Nemotron-3-Nano	3000	0.265	0.277	2206	0.15	0.72	0.13	0.370
Gemma-4-31B, Assamese	251	0.737	0.896	66	0.73	0.26	0.01	0.043
Gemma-4-31B, Bengali	250	0.784	0.912	54	0.72	0.20	0.07	0.036
Gemma-4-31B, Hindi	250	0.904	0.940	24	0.58	0.29	0.12	0.022
Nemotron-3-Nano, Assamese	251	0.171	0.171	208	0.11	0.80	0.09	0.535
Nemotron-3-Nano, Bengali	250	0.276	0.384	181	0.28	0.65	0.07	0.348
Nemotron-3-Nano, Hindi	250	0.636	0.600	91	0.41	0.51	0.09	0.289

Table 30: Separating execution from comprehension (2,943 paired items, reasoning off; * interval excludes zero). *calc*: natural-language reasoning in the target language with every arithmetic step delegated to a calculator we run; *program*: a straight-line program that we execute; *program, English*: the identical program prompt on the oracle English text, so the output register is held fixed and only the input language changes; *English*: natural-language reasoning on the oracle English text. Removing arithmetic (calc, program) helps Gemma-4-31B and Sarvam-105B and no weak model, and giving Gemma-4-31B English input on top of a program adds nothing, consistent with an execution component in its residual gap. For Gnani Evon 3.3 and Nemotron the program register alone does not help and English input does, in both output registers: this is consistent with a comprehension component, and holding the program register the same in both arms of that contrast reduces, without excluding, weak code generation as its explanation. Restricting any contrast to census-valid items moves it by at most 3.1 points.

	accuracy					execution removed		English input	
Model	zero-shot	calc	program	program, En.	English	calc − zero-shot	program − zero-shot	program: En. − target	prose: En. − target
Gemma-4-31B	0.862	0.896	0.913	0.902	0.878	+0.034 [+0.010, +0.064]*	+0.051 [+0.025, +0.081]*	-0.011 [-0.022, -0.001]*	+0.016 [-0.006, +0.044]
Sarvam-105B	0.849	0.873	0.867	0.902	0.875	+0.023 [+0.006, +0.038]*	+0.018 [-0.004, +0.037]	+0.035 [+0.015, +0.060]*	+0.025 [+0.007, +0.048]*
Gemma-4-26B	0.839	0.857	0.865	0.885	0.819	+0.018 [-0.008, +0.049]	+0.026 [-0.003, +0.062]	+0.020 [+0.007, +0.032]*	-0.020 [-0.046, +0.008]
Qwen3.8-27B	0.819	0.823	0.802	0.868	0.882	+0.003 [-0.014, +0.021]	-0.018 [-0.039, +0.003]	+0.066 [+0.027, +0.109]*	+0.063 [+0.032, +0.097]*
Sarvam-30B	0.809	0.782	0.815	0.890	0.877	-0.027 [-0.042, -0.014]*	+0.005 [-0.005, +0.015]	+0.075 [+0.029, +0.145]*	+0.067 [+0.028, +0.127]*
Gnani Evon 3.3	0.647	0.545	0.590	0.715	0.818	-0.102 [-0.138, -0.066]*	-0.057 [-0.095, -0.013]*	+0.126 [+0.066, +0.193]*	+0.171 [+0.110, +0.267]*
Nemotron-3-Nano	0.265	0.279	0.277	0.693	0.733	+0.015 [-0.022, +0.059]	+0.013 [-0.013, +0.043]	+0.415 [+0.337, +0.496]*	+0.468 [+0.380, +0.555]*
pooled, seven models	0.727	0.722	0.733	0.836	0.840	-0.005 [-0.020, +0.009]	+0.005 [-0.008, +0.018]	+0.104 [+0.077, +0.129]*	+0.113 [+0.086, +0.141]*

Table 31: Gemma-4 scale ladder, thinking off, full test split (difference against the rung's own zero-shot).

arm	E4B	12B	26B-A4B	31B
zero-shot	0.685	0.731	0.843	0.871
English instruction	0.712 (+0.027)	0.772 (+0.041)	0.845 (+0.002)	0.897 (+0.026)
all digits ASCII	0.716 (+0.031)	0.775 (+0.044)	0.862 (+0.019)	0.883 (+0.012)
Indic few-shot, k=5	0.675 (−0.010)	0.763 (+0.032)	0.870 (+0.027)	0.895 (+0.024)
language spread, zero-shot	0.159	0.201	0.133	0.148

E.5 Gaps 4–6: sampling headroom, language of thought, robustness

Table 32: Pass@k on the 3,000-item subsample, thinking on, vendor decoding, 32 samples. Gemma-4-26B-A4B and Qwen3.8-27B were not run.

model	pass@1	@2	@4	@8	@16	@32	maj@32	never solved
Gemma-4-31B	0.915	0.933	0.944	0.952	0.958	0.962	0.924	0.038
Sarvam-105B	0.852	0.908	0.934	0.950	0.961	0.969	0.908	0.031
Evon 3.3	0.836	0.897	0.925	0.941	0.952	0.959	0.895	0.041
Sarvam-30B	0.805	0.871	0.905	0.926	0.939	0.950	0.876	0.050
Nemotron-3-Nano	0.569	0.651	0.710	0.754	0.790	0.822	0.655	0.178

Table 33: Forced language of thought by prefill, 3,000 items, thinking on, greedy. Native share (the script of the *question's* language) and Latin share are measured on what the model wrote *after* the prefilled sentence. The forced-Hindi arm drops the 250 Hindi items ($n = 2{,}750$); there, characters that are neither native nor Latin are Devanagari. The other four models were not run. *Cap-hit* is the share of generations that reach the token cap; the truncated share, which also requires that no answer parses, is 0.0–3.9 points lower in the forced arms (13.4% and 36.2% for Evon 3.3, 11.8% and 26.9% for Nemotron, English and native).

model	arm	acc.	cap-hit	native script in thought (mean share / thoughts >50%)	Latin share	mean tokens
Evon 3.3	free choice	0.817	6.5%	0.007 / 0.4%	0.993	2,969
	forced English	0.751	13.8%	0.007 / 0.2%	0.993	4,864
	forced native	0.547	37.1%	0.956 / 99.2%	0.044	11,432
	forced Hindi	0.536	36.2%	0.898 / 93.7%	0.048	11,216
Nemotron-3-Nano	free choice	0.579	15.6%	0.095 / 0.0%	0.905	7,150
	forced English	0.465	15.7%	0.065 / 0.2%	0.934	7,050
	forced native	0.286	28.1%	0.897 / 93.2%	0.101	9,082
	forced Hindi	0.308	12.3%	0.174 / 16.7%	0.117	4,428
Qwen3.8-27B	free choice	0.915	1.1%	0.153 / 0.7%	0.847	2,576
	forced English	0.881	2.6%	0.125 / 0.0%	0.875	2,501
	forced native	0.841	6.7%	0.877 / 93.4%	0.123	3,207
	forced Hindi	0.807	7.6%	0.191 / 13.0%	0.136	3,331

Table 34: Forced language of thought, split by completion (3,000 items, one prefilled sentence in English or in the target language; same token cap in both arms). *Cap-hit*: the response reached the token cap; 1–25% of cap-hit responses still contain a parseable answer and are scored on it, so this is not the *truncated* share of Appendix E, which also requires that no answer parses. *Completed*: accuracy among responses that did not reach the cap. The last column is a descriptive paired contrast on items that *both* arms complete. Completion depends on the intervention, so that contrast does not identify a truncation-free population effect or a formal bound, and the all-items contrast is the observed effect at this budget, not a bound on an intrinsic one. For Gnani Evon 3.3 and Qwen3.8 most of the all-items loss coincides with cap-hits; for Nemotron it does not.

	all items			**cap-hit**		**completed**		**both complete**	
Model	en	native	Δ	en	native	en	native	n	Δ
Gnani Evon 3.3	0.751	0.547	-0.204 [-0.221, -0.186]	0.138	0.371	0.866	0.859	1805	-0.042 [-0.056, -0.027]
Qwen3.8-27B	0.881	0.841	-0.040 [-0.053, -0.027]	0.026	0.067	0.905	0.898	2739	-0.011 [-0.021, -0.001]
Nemotron-3-Nano	0.465	0.286	-0.178 [-0.196, -0.161]	0.157	0.281	0.529	0.394	1916	-0.165 [-0.188, -0.144]

Table 35: Language spread by reasoning regime (3,000-item subset, 250 per language, greedy decoding). *Spread* is best-minus-worst language accuracy in points, *lowest* the worst language, *cap-hit* the share of generations that reach the token cap. Reasoning narrows the spread for four models and widens it for Nemotron; for the two Gemma models the reasoning-on cells are greedy long-form runs that loop (§6), so their narrower spread comes with lower accuracy. The two Sarvam models are decoded with sampling only and are not in this comparison.

	reasoning off				**reasoning on**			
Model	acc.	spread	lowest	cap-hit %	acc.	spread	lowest	cap-hit %
Gemma-4-31B	0.863	17.9	Assamese	0.3	0.808	8.0	Kannada	12.3
Gemma-4-26B	0.840	16.7	Assamese	1.8	0.625	14.4	Kannada	37.2
Qwen3.8-27B	0.818	21.7	Tamil	1.4	0.911	8.1	Tamil	1.2
Gnani Evon 3.3	0.646	52.8	Odia	7.3	0.813	25.2	Odia	6.1
Nemotron-3-Nano	0.265	60.8	Odia	9.4	0.569	85.6	Odia	16.7

Table 36: Input perturbations on the accepted subsample items (paired difference against the same model's zero-shot, same mode, same items). Gemma-4-26B-A4B's thinking-on rows are dominated by its greedy loop (34% truncated at baseline, 36–49% in the arms) and must not be read as perturbation effects; Gemma-4-31B's are affected to a lesser degree (12% truncated at baseline, 11–18% in the arms, 17.6% under oracle English, which is why English input *costs* it points with thinking on).

mode	arm	G-26B-A4B	G-31B	Nemotron	Qwen3.8	Evon 3.3	S-105B	S-30B
off	oracle English	0.819 (−0.020)†	0.878 (+0.016)†	0.733 (+0.468)†	0.882 (+0.063)†	0.818 (+0.171)†	0.875 (+0.025)†	0.877 (+0.067)†
	code-mixed	0.843 (+0.012)	0.864 (+0.009)	0.376 (+0.127)†	0.837 (+0.021)†	0.674 (+0.046)†	0.854 (+0.018)†	0.855 (+0.045)†
	distractor	0.779 (−0.060)†	0.812 (−0.047)†	0.202 (−0.061)†	0.751 (−0.066)†	0.554 (−0.092)†	0.768 (−0.078)†	0.722 (−0.086)†
	paraphrase	0.812 (−0.027)†	0.851 (−0.011)†	0.298 (+0.028)†	0.815 (−0.004)	0.646 (−0.008)	0.838 (−0.011)	0.808 (+0.002)
	romanised	0.705 (−0.135)†	0.826 (−0.037)†	0.102 (−0.164)†	0.606 (−0.212)†	0.448 (−0.200)†	0.785 (−0.064)†	0.685 (−0.125)†
on	zero-shot	0.625	0.808	0.569	0.911	0.813	0.862	0.820
	oracle English	0.479 (−0.145)†	0.750 (−0.058)†	0.887 (+0.316)†	0.917 (+0.006)	0.871 (+0.057)†	0.888 (+0.025)†	0.885 (+0.063)†
	code-mixed	0.597 (−0.006)	0.797 (+0.007)	0.716 (+0.174)†	0.919 (+0.016)†	0.816 (+0.016)	0.864 (+0.007)	0.853 (+0.038)†
	distractor	0.547 (−0.070)†	0.769 (−0.033)†	0.500 (−0.070)†	0.842 (−0.070)†	0.739 (−0.070)†	0.777 (−0.086)†	0.738 (−0.082)†
	paraphrase	0.596 (−0.023)†	0.785 (−0.018)†	0.606 (+0.025)†	0.898 (−0.015)†	0.801 (−0.017)†	0.848 (−0.011)	0.818 (−0.001)
	romanised	0.479 (−0.145)†	0.786 (−0.021)†	0.291 (−0.280)†	0.766 (−0.145)†	0.613 (−0.200)†	0.790 (−0.072)†	0.686 (−0.135)†

Table 37: Few-shot accuracy, thinking off, full test split (paired difference against zero-shot). "English" exemplars have English solutions; "Indic" exemplars are in the item's language; the cross-lingual row uses Hindi exemplars (Bengali for Hindi items). *: the 95% paired interval excludes zero; numeric intervals are in the released results. Differences of about one point are at the resolution limit of this comparison ($+0.009$ is starred for one model and -0.011 is not for another).

exemplars	k	G-26B-A4B	G-31B	Nemotron	Qwen3.8	Evon 3.3	S-105B	S-30B
none	0	0.843	0.871	0.262	0.828	0.661	0.865	0.823
English	1	0.851 (+0.009)*	0.899 (+0.028)*	0.380 (+0.118)*	0.841 (+0.013)*	0.752 (+0.091)*	0.820 (−0.045)*	0.801 (−0.023)*
	3	0.856 (+0.013)*	0.901 (+0.031)*	0.375 (+0.114)*	0.823 (−0.005)	0.736 (+0.074)*	0.814 (−0.051)*	0.799 (−0.025)*
	5	0.860 (+0.017)*	0.899 (+0.029)*	0.372 (+0.111)*	0.818 (−0.010)*	0.710 (+0.049)*	0.807 (−0.058)*	0.795 (−0.028)*
	8	0.869 (+0.026)*	0.903 (+0.032)*	0.365 (+0.103)*	0.811 (−0.017)*	0.693 (+0.032)*	0.764 (−0.101)*	0.793 (−0.031)*
Indic	1	0.860 (+0.017)*	0.884 (+0.013)*	0.301 (+0.039)*	0.816 (−0.012)*	0.680 (+0.019)*	0.804 (−0.061)*	0.817 (−0.007)
	3	0.867 (+0.024)*	0.897 (+0.026)*	0.290 (+0.028)*	0.800 (−0.028)*	0.672 (+0.011)*	0.810 (−0.055)*	0.810 (−0.014)*
	5	0.870 (+0.028)*	0.895 (+0.024)*	0.291 (+0.030)*	0.785 (−0.043)*	0.654 (−0.007)	0.806 (−0.059)*	0.803 (−0.020)*
	8	0.873 (+0.030)*	0.895 (+0.024)*	0.290 (+0.029)*	0.784 (−0.044)*	0.634 (−0.027)*	0.773 (−0.092)*	0.800 (−0.024)*
cross-lingual	5	0.844 (+0.001)	0.880 (+0.009)*	0.291 (+0.029)*	0.784 (−0.044)*	0.651 (−0.011)	0.762 (−0.103)*	0.728 (−0.095)*

F Re-minting and source-identical operands

Stage 1 draws every operand afresh from the range Stage 0 attached to it. For some operands Stage 0 emits a degenerate range $[v, v]$ equal to the source value (a constant the problem depends on, or a value the extractor would not vary), and the sampler then returns v. An item is *source-identical* when every released operand equals its source's: 28.6% of the corpus and 2,714 of 8,860 public test items (30.6%). The other items were already resampled once at generation, which makes them a control stratum for re-minting. On all 1,280 eligible contracts (628 source-identical) we resample operands at the same magnitude and keep the question text, three independent draws, for seven models. Table 38 gives the paired drop by stratum with language-clustered intervals, and the zero-shot accuracy of each model on the two strata of the full split.

Table 38: Re-minting at full power: 1,280 eligible contracts, all seven models, reasoning off, the released item against the *mean of three* independent same-magnitude operand draws with the question text kept (paired per item; language-clustered bootstrap, 2,000 resamples; * interval excludes zero). The drop is split by whether the released operands equal the source problem's (628 items) or had already been resampled by Stage 1 (652 items). The already-resampled stratum is a control: it receives the identical re-minting procedure and shows no detectable average drop for any model, The strata are not randomly assigned, so an accuracy-stable control does not exclude differential changes in validity or difficulty (Table 39). The whole effect sits on the source-identical items, and the difference between strata excludes zero for six of seven models. Restricting to the 684 items whose gold stays integral gives a pooled drop of -0.070 $[-0.087, -0.054]$, so the effect is not a by-product of non-integer answers; a larger-magnitude draw costs -0.102 and a decimal draw -0.088. The last two columns give full-split zero-shot accuracy by stratum (2,714 source-identical and 6,146 already-resampled test items): the three tiers of Table 3 hold in both, while inside the leading tier the two Gemma models gain about ten points from source-identical operands and the two Sarvam models under five. The stratum flag ships per item; the already-resampled column is the ranking we recommend when source familiarity is a concern.

		re-mint drop (mean of three draws − released)				**zero-shot, full split**	
Model	released	all eligible	source-identical	already resampled	difference	source-id.	resampled
Gemma-4-31B	0.920	-0.063 [-0.082, -0.046]*	-0.109 [-0.147, -0.077]*	-0.018 [-0.041, +0.006]	-0.091 [-0.141, -0.049]*	0.941	0.840
Sarvam-105B	0.900	-0.049 [-0.069, -0.032]*	-0.091 [-0.119, -0.066]*	-0.008 [-0.033, +0.021]	-0.083 [-0.125, -0.046]*	0.895	0.852
Gemma-4-26B	0.888	-0.049 [-0.068, -0.033]*	-0.097 [-0.129, -0.068]*	-0.003 [-0.017, +0.011]	-0.094 [-0.126, -0.065]*	0.913	0.812
Qwen3.8-27B	0.864	-0.045 [-0.058, -0.031]*	-0.098 [-0.118, -0.076]*	+0.007 [-0.012, +0.023]	-0.105 [-0.132, -0.078]*	0.853	0.817
Sarvam-30B	0.867	-0.046 [-0.066, -0.027]*	-0.075 [-0.098, -0.049]*	-0.018 [-0.045, +0.005]	-0.056 [-0.089, -0.025]*	0.828	0.822
Gnani Evon 3.3	0.759	-0.056 [-0.068, -0.044]*	-0.097 [-0.130, -0.063]*	-0.015 [-0.037, +0.005]	-0.082 [-0.129, -0.030]*	0.721	0.635
Nemotron-3-Nano	0.362	-0.007 [-0.020, +0.007]	-0.024 [-0.057, +0.010]	+0.009 [-0.027, +0.040]	-0.033 [-0.091, +0.031]	0.311	0.240
pooled, seven models	0.794	-0.045 [-0.056, -0.036]*	-0.085 [-0.102, -0.068]*	-0.007 [-0.018, +0.004]	-0.078 [-0.100, -0.058]*		

Table 39 is the complementary analysis: it restricts to pairs a plausibility judge called well posed in both versions and bootstraps over items, and at its sample size it does not separate the strata. The judge is Qwen3-235B with reasoning on at the vendor's sampling setting (T=0.6, top-p 0.95, top-k 20, one seeded pass; prompt in the release). It sees 300 source-identical and 300 already-resampled contracts, each as the released question and as its first re-mint without being told which, and answers three yes/no fields (well posed, plausible, natural numbers); all 1,200 verdicts parsed. It is a single pass with no repeat-judging or human agreement estimate. A pair is *clean* when both versions are well posed and plausible: 181 of 300 source-identical and 176 of 300 control pairs. Exclusions are not neutral: contracts of complexity level 3 or above are 41% of excluded against 31% of retained source-identical pairs (18% against 9% in the control), and a re-mint alone fails on 66 source-identical against 40 control pairs, so the clean subset is easier and the restriction removes more source-identical re-mints. The two analyses answer different questions. The full evaluation shows greater re-minting sensitivity among source-identical items; the control stratum shows no detectable average drop (-0.007 $[-0.018, +0.004]$). Because the strata were not randomly assigned, differential changes in validity or difficulty remain possible. The judged-clean analysis estimates a reduced source-identical loss (3.6 points) and does not resolve the difference between strata (-0.022 $[-0.064, +0.020]$).

Alternative explanations. *Rounder numbers.* Re-instantiated items can be harder simply because human-chosen operands are rounder (Długosz et al., 2026). Same-magnitude re-mints match the released questions on significant digits, and a per-item regression of the drop on the change in question and answer digits leaves about 80% of it at equal roundness. *Arithmetic load.* On the $\approx$240 source-identical items whose re-mint is no harder in intermediates, question digits or answer digits the drop is still 5–12 points for the three strongest greedy models. Neither control turns the residual into a

Table 39: Pooled re-mint drop (first same-magnitude draw against the released item, seven models; truncated responses excluded) on the 600 contracts whose two versions were both rated by the plausibility judge (Qwen3-235B, reasoning on, one pass, blind to which version it sees; 181 of 300 source-identical and 176 of 300 control pairs are clean in both versions). n is model–item pairs / distinct items. Intervals resample *items*, carrying all of an item's model rows together, because the seven models answer the same items. The *difference* rows subtract the already-resampled control. On pairs judged well posed and plausible in both versions the source-identical drop falls to 3.6 points and remains distinguishable from zero, and its difference from the control is not resolved: conditioning on judged validity lowers the estimate as well as the sample, so this table qualifies the full-sample contrast of Table 38.

	all judged pairs		**judged clean in both versions**	
Slice	n	Δ	n	Δ
source-identical, reasoning off	2021 / 300	-0.077 [-0.108, -0.047]	1228 / 181	-0.036 [-0.069, -0.003]
already resampled (control), reasoning off	2048 / 300	-0.011 [-0.033, +0.012]	1206 / 176	-0.014 [-0.042, +0.013]
difference, reasoning off		-0.066 [-0.104, -0.028]		-0.022 [-0.064, +0.020]
source-identical, reasoning on	1774 / 300	-0.067 [-0.096, -0.038]	1086 / 181	-0.036 [-0.070, -0.003]
already resampled (control), reasoning on	1769 / 300	-0.018 [-0.039, +0.003]	1068 / 176	-0.009 [-0.035, +0.016]
difference, reasoning on		-0.048 [-0.085, -0.013]		-0.026 [-0.070, +0.014]

measurement of contamination: that source authors pick numbers that are friendlier in ways these covariates miss is not excluded. *One factor at a time.* Holding the numbers fixed and changing only the grounding (the six variants of §C.3) moves accuracy by $+0.2$ to -1.8 points with no interval excluding zero; changing only the wording (a fresh re-voicing of the same translation) costs 0.7 points for Gemma-4-31B (n.s.) and 2.3 [0.2, 4.5] for Qwen3-235B (Table 40). Re-minted versions should therefore be compared at the split level, never item by item.

Table 40: Accuracy change from the released text when one factor is re-minted (paired, 95% bootstrap within language). $^{\equiv}$: equivalent within ± 2 points (90% interval inside the margin).

Factor	**Re-mint**	**Gemma-4-31B**	**Qwen3-235B**
grounding	neutral (no regional detail)	0.000 [−0.019, +0.019]$^{\equiv}$	−0.007 [−0.031, +0.019]
	names only	−0.003 [−0.024, +0.017]	+0.002 [−0.022, +0.024]
	places only	−0.003 [−0.025, +0.019]	−0.007 [−0.036, +0.020]
	places removed	0.000 [−0.015, +0.015]$^{\equiv}$	−0.015 [−0.035, +0.007]
	unit explained	0.000 [−0.018, +0.017]$^{\equiv}$	−0.018 [−0.038, +0.002]
	moved to a distant state	+0.002 [−0.012, +0.015]$^{\equiv}$	−0.015 [−0.038, +0.007]
wording	re-voicing replay (mean of 3 seeds)	−0.007 [−0.027, +0.012]	−0.023 [−0.045, −0.002]

Templates and near-duplicate sources. Splits are disjoint by source problem, but 22.8% of test items use a formula template that never occurs in training, and 4,651 of 6,345 test sources have a training-side near-twin *in English* (word trigrams, numbers masked, Jaccard ≥ 0.8), a property of AQuA-RAT and MathQA and not of our generation. Items with a near-twin source are never easier (gaps of 0.0 to -6.0 points; at most $+0.8$ after reweighting to a common complexity and language mix), and unseen templates separate models without favouring them (-1.1 to -3.7 points after reweighting for the models with a negative raw gap; both Gemma models are better on unseen templates; Table 41). None of these models was trained on the training split, so these are coverage gaps, not leakage. That changes for anyone who *does* train on the released training split: near-twin sources then bear on test performance. We therefore release the near-twin flag, and studies that train on BharatMath should report the 2,270 test items with no training-side near-twin separately.

Table 41: Zero-shot accuracy (public test, thinking off) by overlap with the training split. Gaps with 95% bootstrap intervals. "Common" templates occur at least ten times in training; the 2,059 test items whose template occurs one to nine times are in neither column. *adj.*: the comparison group reweighted to the complexity × language mix of the first group.

	template unseen (2,022) vs **common** (4,779)			**near-duplicate source** (6,590) vs **clean** (2,270)		
Model	unseen	gap	common adj.	near-dup.	gap	clean adj.
Gemma-4-31B	.898	+0.030 [+0.013, +0.047]	.837	.864	−0.026 [−0.041, −0.011]	.887
Gemma-4-26B	.850	+0.001 [−0.019, +0.019]	.828	.836	−0.027 [−0.044, −0.011]	.860
Qwen3.8-27B	.803	−0.048 [−0.067, −0.028]	.840	.828	−0.000 [−0.018, +0.018]	.820
Sarvam-30B	.808	−0.037 [−0.058, −0.017]	.837	.819	−0.017 [−0.034, +0.001]	.829
Gnani Evon 3.3	.631	−0.065 [−0.090, −0.040]	.642	.646	−0.060 [−0.082, −0.038]	.697
Nemotron-3-Nano	.203	−0.098 [−0.120, −0.075]	.221	.249	−0.051 [−0.073, −0.030]	.282

G Release

The full corpus contains 88,066 items. The public release includes 70,336 training and 8,860 public-test items, totalling **79,196**; the remaining 8,870 held-out items stay private, since publishing that partition, whether as a corpus file or a split map, would remove its private status. The release also includes the three Context Bank workbooks, all prompts and demonstrations, per-item lineage and quality flags (constant formulas, source-identical operands, float-printed integers, the census validity label of the 8,778 labelled public test items, flagged blueprints, near-twin sources), per-item predictions for every arm reported, and the scripts that regenerate every table. Additional arms are documented with their predictions in the supplementary experiment record.