ANCHOR: CROSS-LINGUAL REASONING WITH ANCHORED REINFORCEMENT LEARNING UNDER ACCURACY-LANGUAGE TRADE-OFF

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought LLMs support multilingual reasoning, yet on low-resource language questions, they often rely on high-resource languages such as English for accuracy. Forcing reasoning in the question language improves readability for question-language speakers but may compromise accuracy. To address this trade-off, we propose Anchor[1], a representation anchoring and behavioral framework for cross-lingual reasoning. We structure the response into three segments: a multilingual reasoning segment, a question-language public explanation, and a verifiable answer. At the representational level, layer-wise probing reveals a recovery pattern in question language signals, based on which we introduce a one-sided anchoring loss to preserve question language routing; at the behavioral level, we design a multi-dimensional reward with two language-gated rewards to shape language choices under answer correctness and format validity. Both levels are jointly optimized in a DAPO-based policy optimization process. Experiments show Anchor enables flexible cross-lingual reasoning on multiple benchmarks. On MGSM, Anchor reaches 86.07% average accuracy across 11 languages, delivering 1.90% and 1.94% gains over the baseline for 4 low-resource and 7 high-resource languages respectively, alongside 74.2% public-segment question-language alignment. On MMLU-ProX-Lite and Global-MMLU-Lite, Anchor boosts performance on 41 of 52 languages and modest alignment improvements.

1 INTRODUCTION

Long Chain-of-Thought(CoT) is an effective paradigm for reasoning tasks, yet its advances are dominated by high-resource languages like English(Wei et al., 2022; Muennighoff et al., 2025; Yong et al., 2025). Given low-resource inputs (Swahili or Tibetan), models often reason in high-resource languages even when the question itself is written in the question language(Yong et al., 2025; Saji et al., 2026; Park et al., 2026). Though such outputs can be correct, they hinder reading and verification by question language users(Saji et al., 2026). Meanwhile, enforcing question language reasoning throughout degrades accuracy on low-resource languages and complex tasks(Saji et al., 2026; Son et al., 2026; Wang et al., 2025). We thus investigate flexible multilingual reasoning.

This goal poses two model-level challenges. (1), language-related representation layer: representations of different languages occupy partially differentiated regions of the model's hidden-state space(Wendler et al., 2024). These layer-specific patterns elude direct control by output rewards. (2), reward-driven behavioral layer: existing Reinforcement Learning (RL) rewards mainly target answer correctness and ignore language choice during generation. Thus, models lack an incentive to retain the question language, favoring high-resource languages for reasoning(Park et al., 2026).

We propose **Anchor**, a behavioral and representation anchoring framework for cross-lingual long CoT reasoning. We structure each output into three segments: a reasoning segment that allows flexible mixing of the question language and high-resource languages such as English, a public explanation segment restricted to the question language, and an answer segment with constrained syntax for automatic verification. To achieve this output schema, Anchor shapes language use at both the

[1]Anonymous repo: `https://anonymous.4open.science/r/Anchor-CDE1/`. Code will be released upon acceptance.

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.

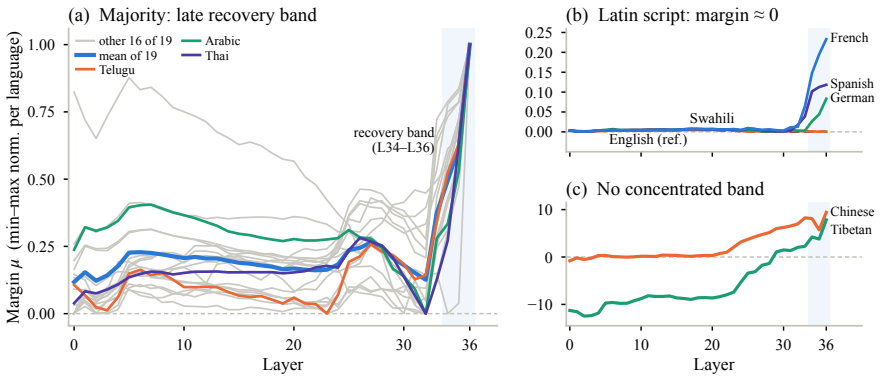

Figure 1: Layer-wise vocabulary margins via logit lens for 26 probed languages. **(a)** 19 of 21 non-Latin-script languages (min–max normalized; mean in blue; Telugu, Arabic, Thai highlighted). **(b)** Five Latin-script evaluation languages (raw scale). **(c)** Chinese and Tibetan (raw scale).

representation and behavioral levels. At the representation level, to examine internal language routing, we perform a layer-wise logit-lens analysis(nostalgebraist, 2020) on the initial model[2] across 26 languages(Figure 1). We observe three distinct patterns. Nineteen of the 21 non-Latin languages show margin decay in mid-late layers followed by recovery confined to the final layers, termed the *recovery band*. The five Latin-script languages maintain margins 1–2 orders of magnitude smaller than non-Latin counterparts, showing only minor band upticks with no drop–recover behavior. Chinese and Tibetan exhibit monotonic margin growth with depth and lack this near-output recovery band. Motivated by this observation, we introduce a one-sided anchoring loss, constraining vocabulary margins to the initial model's lower bound in the recovery band, to preserve question-language routing during RL. The constraint applies to 21 non-Latin-script languages among the 50 training languages for which the vocabulary margin is informative, while Latin-script languages are shaped behaviorally. At the behavioral level, we design a multidimensional reward function: language rewards based on fastText identification cover 44 of the 50 training languages and are gated by correctness and format validity, separately shaping public-segment alignment and reasoning-segment language mixing. Quality penalties are applied independently. We further use a discrete group score and filter groups with insufficient language-behavior variation to preserve the relative differences. Both levels are jointly optimized within Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO)(Yu et al., 2026).

2 METHOD

Anchor formulates cross-lingual long-CoT as structured generation with three response segments. As shown in Figure 2, it jointly constrains question-language routing in the recovery band and language use throughout the output within a single policy optimization process of DAPO.

2.1 PROBLEM FORMULATION

Given a question q written in its question language g, the model generates $y = (r, e, \hat{a})$, consisting of a reasoning segment r, a public explanation e, and a canonical answer $\hat{a}$. The reasoning segment allows flexible multilingual use, while e is restricted to g. The answer $\hat{a}$ is a number or option letter for automatic grading. We call this convention the output contract.

2.2 REPRESENTATION LEVEL: RECOVERY BAND ANCHORING

The representational level preserves question language routing during reinforcement learning. We first identify a recovery band where question-language signals recover, and then anchor the vocabulary margin to an initial-model lower bound within this band during reinforcement learning.

[2]The initial model mentioned in the INTRODUCTION and METHOD sections is Anchor's backbone Qwen3-8B-Instruct. We discussed the performance of other initial models in the EXPERIMENTS section.

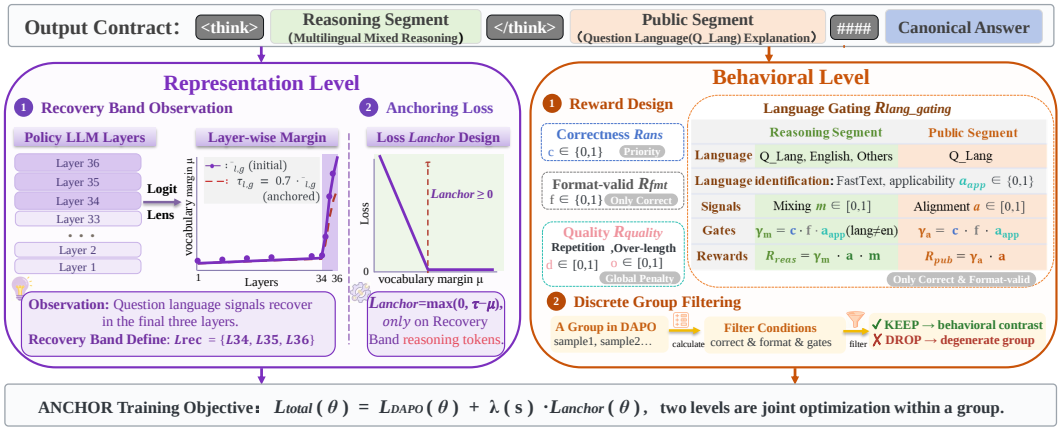

Figure 2: Overview of Anchor. Top: the three-segment output contract. Middle: the behavioral constraints and the representational constraint. Bottom: the joint objective.

Recovery-Band Observation. We perform a layer-wise logit-lens analysis on the initial model. Given a question, we project the hidden states at token positions in its second half into the vocabulary space through the output head and compute the vocabulary margin, we define vocabulary margin as:

$$\mu_l(t) = \mathrm{LSE}_{v\in V_g} z_l(t)_v - \mathrm{LSE}_{v\in V_{\mathrm{lat}}} z_l(t)_v, \tag{1}$$

where LSE denotes the log-sum-exp, V_g is the question-language token set and V_{lat} is the Latin-script token set. This margin is the log ratio of their probability masses and a positive margin means the question-language vocabulary is relatively dominant. Across most of the 21 Anchor-eligible languages, the margin weakens to near zero or below in intermediate layers and recovers mainly in the final three layers. We refer to these layers as the recovery band L_{rec}.

anchoring Loss. Based on the recovery band observation, we design a one-sided anchoring loss that uses the initial model's vocabulary margin as a local lower bound, preserving question language routing while allowing flexible cross-lingual reasoning. The constraint is applied only within the recovery band and the latter half of the reasoning segment, we define the anchoring loss $\mathcal{L}_{Anchor}$ as:

$$\mathcal{L}_{Anchor}(\theta) = \frac{1}{|T_{\mathrm{Anchor}}|\,|L_{\mathrm{rec}}|} \sum_{t\in T_{\mathrm{Anchor}}} \sum_{l\in L_{\mathrm{rec}}} \max\big(0,\ \tau_{l,g} - \mu_l^\theta(t)\big), \qquad \tau_{l,g} = \alpha\,\bar{\mu}_{l,g},\ \ \alpha = 0.7, \tag{2}$$

where T_{Anchor} contains the latter 50% of reasoning segment tokens. $\mu_l^\theta(t)$ is the policy margin, and $\bar{\mu}_{l,g}$ is the corresponding initial-model margin for question language. The hinge term penalizes only margins below the bound; with $\alpha = 0.7$, the bound moves 30% toward zero in log-ratio space, preserving part of the initial preference when $\bar{\mu}_{l,g} > 0$ while partially reducing its disadvantage when $\bar{\mu}_{l,g} < 0$. During training, final-layer logits are used directly and intermediate states are projected through the logit lens. The loss applies to the 21 Anchor-eligible languages (the non-Latin-script) whose token sets are separable from V_{lat}; for Latin-script languages the margin is undefined, and they rely on the behavioral level. Probe construction, per-language recovery-band margins, and anchoring bounds are provided in Appendix B.

2.3 Behavioral Level: Language Reward and Group Filtering

The behavioral level shapes language use in generated responses via two mechanisms: reward design for scoring individual outputs, and group filtering for selecting valid groups for optimization.

Reward Design. We design a multi-dimensional reward with correctness-aware language shaping. The reward jointly captures answer correctness, output format, language use in the public and reasoning segments, and generation quality. We decompose the reward into four components:

$$R_{\mathrm{total}} = R_{\mathrm{ans}} + R_{\mathrm{fmt}} + R_{\mathrm{lang_gate}} - R_{\mathrm{quality}}; \tag{3}$$

where R_{ans} and R_{fmt} evaluate the canonical answer and output contract; $R_{\mathrm{lang_gate}}$ shapes language use separately in public and reasoning segments; R_{quality} penalizes repetition and excessive length.

(1) Answer and Format Rewards. Answer correctness is the primary objective, while format validity is rewarded only for correct answers. The answer reward is $R_{\text{ans}} = \beta_c c$, and the format reward is $R_{\text{fmt}} = c\beta_f f$, where $c \in \{0, 1\}$ indicates whether the answer is correct and $f \in \{0, 1\}$ indicates whether the response can be fully parsed into the three segments. We set $\beta_c = 2.0$ and $\beta_f = 0.1$.

(2) Language Gating Reward. Language gating reward $R_{\text{lang_gate}}$ is applied only to correct and format-valid responses. It comprises two rewards, R_{pub} for question language g use in the public segment and R_{reas} for question multi-language mixing in the reasoning segment, with $R_{\text{lang}_{\text{g}}\text{ate}} = R_{\text{pub}} + R_{\text{reas}}$. We use label probabilities from the fastText model (Joulin et al., 2016b;a) [3] to discriminate languages. Limited by fastText's language coverage, $a_{\text{app}} \in \{0, 1\}$ indicates whether the question language g is supported; $R_{\text{lang_gate}}$ are not applied to unsupported languages. Among the 50 training languages, 44 are supported; 6 exceptions only receive R_{ans}, R_{fmt}, and R_{quality}.

For the public segment, $a \in [0, 1]$ denotes the average fastText probability assigned to the question language. We define the gate as $\gamma_a = cfa_{\text{app}}$ and the corresponding reward as $R_{\text{pub}} = \beta_a \gamma_a a$, where β_a=0.15. R_{pub} increases with question-language alignment in the public explanation.

For the reasoning segment, p_g, p_{en}, and p'_o denote the average fastText probabilities of the question language, English, and all other languages. We define the reasoning segment mixing degree as $m = \min(\frac{p_g}{\rho}, 1) \min(\frac{p_{en}}{\rho}, 1)(1 - p'_o)$, where ρ=0.2 is the saturation threshold for language participation. The gate and reward are $\gamma_m = cfa_{\text{app}} \mathbf{1}[g \neq \text{en}]$ and $R_{\text{reas}} = \beta_m \gamma_m ma$, where β_m=0.10. The product form of m rewards sufficient participation of both the question language and English, while the saturation terms prevent continuous gains from increasing either language's proportion and $(1 - p'_o)$ discourages drift toward unrelated languages. For English questions, R_{reas} is disabled.

(3) Quality Penalties. To suppress repetitive or excessively long generations, we introduce two quality penalties that apply to every generated response. Specifically, $R_{\text{quality}} = \beta_d d + \beta_o o$, where $d \in [0, 1]$ measures repetition and $o \in [0, 1]$ measures overlength. We set $\beta_d = 0.3$ and $\beta_o = 0.2$.

Discrete Group Filtering. DAPO requires meaningful behavioral contrast within each response group, but small continuous reward variations may make a behaviorally degenerate group appear informative. We define a discrete group score s_{grp} from correctness, format validity, public-segment alignment, and reasoning-segment mixing; groups with identical scores are discarded as degenerate.

$$s_{\text{grp}} = 4c + 2cf + \gamma_a \mathbf{1}[a \geq 0.5] + \gamma_m \mathbf{1}[a \geq 0.5] \mathbf{1}[m \geq 0.5] \ \in \{0, 4, 6, 7, 8\}. \tag{4}$$

Implementation details of language identification, and language coverage in Appendix A.

2.4 Unified Objective

Anchor jointly optimizes representation anchoring and behavioral rewards within a single DAPO-based policy optimization process. The anchoring loss serves as an auxiliary objective that constrains the vocabulary margin within the recovery band, while the behavioral rewards shape output behavior through group-relative advantages. The overall objective $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{DAPO}}(\theta) + \lambda(s)\, \mathcal{L}_{Anchor}(\theta). \tag{5}$$

We set $\lambda(s) = 0.01$, linearly warmed up over the first 5% of training steps. For each question, we sample $N = 8$ responses and adopt DAPO's group-normalized advantages, decoupled clipping, token-level loss aggregation, dynamic sampling, and linear overlength penalty. Omitting KL regularization removes the need for a reference model.

3 Experiments

3.1 Experimental Setup

Datasets. (1) Training data. Our 50-language corpus (21 non-Latin, 29 Latin) is aggregated from MGSM8KInstruct(Chen et al., 2023) augmented with our Tibetan subset, Global-MMLU(Singh et al., 2024) and MMLU-ProX(Xuan et al., 2025). All data is split by semantic group (full translations of a single question) to ensure cross-lingual semantic alignment and prevent leakage. We

[3] fastText model: `https://fasttext.cc/docs/en/language-identification.html`

remove all overlapping semantic groups with the three fixed evaluation data, apply targeted low-resource language oversampling, and obtain 15,243 valid samples across 5,454 groups, split into 14,472 training samples (5,168 groups) and 771 dev samples (286 groups), both covering all 50 languages. The training set contains 4,888 numeric questions from extended MGSM8KInstruct, 4,074 single-choice questions from Global-MMLU, and 5,510 single-choice questions from MMLU-ProX. The dev set is used for model selection and early stopping. **(2) Evaluation data.** We evaluate on three publicly multilingual parallel benchmarks: MGSM(Cobbe et al., 2021; Shi et al., 2022) (11 languages, 2750 samples/250 groups, numeric), Global-MMLU-Lite(Singh et al., 2024) (23 languages, 9,199 samples/400 groups, single-choice), and MMLU-ProX-Lite(Xuan et al., 2025) (29 languages, 17,052 samples/588 groups, single-choice). The original MGSM split into high-resource (en, es, fr, de, zh, ja, ru) and low-resource (sw, bn, th, te) groups(Yong et al., 2025), we extend it with our newly added parallel Tibetan samples.

Metrics. We evaluate three output aspects separately: final answer accuracy for single-choice and numeric tasks, public segment language alignment degree A, and reasoning segment language mixing degree M. For the i-th evaluated response in language g, let $a_{g,i}$ and $m_{g,i}$ denote the per-response alignment and mixing scores defined in Section 2.3, and let N_g be the total number of evaluated responses for language g. We compute the fixed-denominator, language-macro averages $A = \frac{1}{N_g}\sum a_{g,i}$ and $M = \frac{1}{N_g}\sum m_{g,i}$. Higher A indicates stronger question-language alignment, while $m_{g,i} > 0$ indicates the valid multilingual reasoning.

Baselines. (1) Raw open-source LLMs: Qwen3-8B-Base, Qwen3-8B-Instruct, Qwen3-14B (Yang et al., 2025), Gemma-3-27B-it(Team et al., 2025), R1-Distill-Qwen-14B(Guo et al., 2025), and Qwen2.5-14B-Instruct and Qwen2.5-32B-Instruct (Qwen et al., 2025). **(2) Task-oriented LLMs**: The six Qwen2.5 and s1 rows are all taken from (Yong et al., 2025). The CoT process of SFT training data is distilled using Qwen3-8B-Instruct. **(3) Closed-source LLMs**: We further conduct a case analysis on (GPT(Singh et al., 2025) , Gemini(Team et al., 2023) and DeepSeek(Guo et al., 2025)).

Implementation details. We implement Anchor in verl (Sheng et al., 2025) and fully train Qwen3-8B-Instruct for one epoch (seed 42); the main text reports the final checkpoint (template = 0, step = 150). Policy optimization uses asymmetric importance-sampling clipping $[1 - 0.2,\ 1 + 0.28]$ with a dual-clipping lower bound of 10, a zero entropy coefficient, and dynamic sampling that selects 32 valid prompts from 96 candidates per round (at most four resampling attempts). The rollout limit is 6,144 tokens, and the overlength penalty uses a soft threshold of 4,096 tokens, a hard limit of 8,192 tokens, and a maximum magnitude of 0.20. The repetition score d is the maximum of character 8-gram and whole-line repetition over the reasoning and public segments; responses with unparseable reasoning segments or fewer than eight reasoning tokens are excluded from the anchoring loss. The initialization study (Figure 3(a)) additionally uses Qwen3-8B-Base and two distilled SFT models, denoted by $R_{\mathrm{SFT}}(lr = 5 \times 10^{-6})$ and $R_{\mathrm{SFT}}(lr = 1 \times 10^{-6})$, where lr denotes the SFT learning rate. Further details on data construction, the Tibetan extension, distilled initializations, language identification, prompts, grading, and decoding settings are provided in Appendix A.

3.2 Main Results

Table 1 reports Anchor's reasoning accuracy and Language behavior on MGSM (Tibetan extension). Appendix C reports the complete MGSM results, including additional decoding settings, ablations, significance tests, and per-language language-behavior statistics.

Anchor vs. Raw open-source LLMs. Among seven LLMs of comparable or larger scale, Anchor achieves the best average accuracy of 86.07% and consistently outperforms its Qwen3-8B-Instruct backbone by 1.93 points. Performance gains hold across resource tiers: 1.94% improvement on 7 high-resource languages and 1.90% improvement on 4 low-resource languages. In terms of language behavior, compared with the backbone that relies on prompts for language shaping, Anchor obtains a substantial gain of 25.2% in language alignment. Both models support multilingual reasoning.

Anchor vs. Task-oriented LLMs. Among six LLMs larger than 8B adopting either only English or question language reasoning, Anchor achieves the highest AVG (86.07%). Compared with the strongest English-CoT baseline s1-14B-extrapolation, Anchor yields a 5.07% AVG gain, with consistent improvements across all 11 languages and the largest gain of 22.8% on Telugu. Against the strongest Native-CoT baseline Qwen2.5-32B-Instruct, Anchor lifts AVG by 8.67% and surpasses it on all 11 languages, with the maximum improvement of 21.2% on Telugu. These baselines do not model mixed reasoning or public-segment language, so their $\bar{A}$ and $\bar{M}$ are not directly comparable.

Table 1: Results on MGSM: per-language accuracy, average accuracy $\overline{AVG}$, alignment $\overline{A}$ across 11 languages, and mixture $\overline{M}$ over 10 non-English languages, plus standalone Tibetan results. Bold marks category-wise best results; green/red cells denote superior/inferior values among top candidate models. †: Baseline results taken from (Yong et al., 2025).

Model	Accuracy of each language											Summary and extension			
	en	es	fr	de	zh	ja	sw	bn	ru	th	te	AVG	$\overline{A}$	$\overline{M}$	bo
Raw open-source LLMs															
Qwen3-14B †	**94.8**	88.4	**87.2**	83.6	85.2	77.6	63.6	85.2	**94.0**	86.8	**80.0**	**84.22**	–	–	–
Qwen3-8B-Base	84.00	66.00	61.20	66.80	66.00	36.40	12.40	27.60	54.80	69.60	7.60	50.22	0.1	0.1	1.20
Qwen3-8B-Instruct	92.40	**90.80**	85.20	**87.6**	**86.40**	**83.6**	57.60	**85.6**	90.00	**88.0**	78.40	84.15	**49.0**	**8.0**	**7.20**
Gemma-3-27B-it †	88.4	84.0	72.4	83.2	84.4	79.2	**78.0**	64.8	83.2	84.4	76.0	79.82	–	–	–
R1-Distill-Qwen-14B †	83.6	80.4	74.4	77.2	79.6	78.4	22.4	66.0	82.4	74.8	22.4	67.42	–	–	–
Qwen2.5-14B-Instruct †	82.0	77.6	67.6	77.6	84.0	70.4	40.4	74.0	76.4	78.8	50.8	70.9	–	–	–
Qwen2.5-32B-Instruct †	84.0	80.0	69.2	79.6	84.0	80.8	54.0	82.0	77.2	84.4	56.4	75.6	–	–	–
Task-oriented LLMs															
Qwen2.5-14B-Instruct (English-CoT) †	87.6	86.0	68.4	75.2	84.4	76.8	45.6	77.2	76.4	79.2	52.0	73.5	–	–	–
Qwen2.5-14B-Instruct (Native-CoT) †	88.0	87.2	68.4	77.2	83.6	76.0	46.8	79.2	75.6	80.4	53.2	74.1	–	–	–
Qwen2.5-32B-Instruct (English-CoT) †	89.6	84.8	66.8	80.4	84.8	85.2	56.8	82.0	77.6	84.8	55.6	77.1	–	–	–
Qwen2.5-32B-Instruct (Native-CoT) †	90.0	87.2	66.8	78.0	88.0	80.8	57.2	82.4	74.0	87.2	59.6	77.4	–	–	–
s1-14B-truncation (English-CoT) †	92.8	88.4	85.2	84.8	86.4	83.6	55.6	82.0	86.8	85.2	59.6	80.9	–	–	–
s1-14B-extrapolation (English-CoT) †	92.4	86.4	83.2	86.8	87.6	83.2	57.2	82.8	88.8	84.8	58.0	81.0	–	–	–
Anchor	**96.00**	**93.20**	**86.4**	**88.8**	**88.4**	**86.0**	**62.0**	**85.2**	**90.8**	**89.2**	**80.8**	**86.07**	**74.2**	**10.0**	**8.40**

Summary. Anchor provides a competitive method for shaping language behavior without sacrificing accuracy. We perform qualitative case analysis on Tibetan, Bengali and Telugu, comparing GPT, Gemini, DeepSeek-V4-Flash and Anchor; examples are in Appendix E.

3.3 Ablation Study

We conduct ablation studies (Table 2) on MGSM along three dimensions to verify the contribution.

Table 2: Ablation results on MGSM (%). Δ values are measured relative to Anchor. Bold denotes column-best results.

Config	Accuracy of each language											Summary and extension						
	en	es	fr	de	zh	ja	sw	bn	ru	th	te	AVG	Δ_{AVG}	$\overline{A}$	$\Delta_{\overline{A}}$	$\overline{M}$	$\Delta_{\overline{M}}$	bo
Anchor	**96.00**	**93.20**	86.40	**88.80**	**88.40**	**86.00**	**62.00**	85.20	90.80	89.20	**80.80**	**86.07**	–	**74.20**	–	10.00	–	8.40
Core Components																		
R1	93.60	92.00	85.20	**88.80**	86.40	84.40	58.40	84.80	**91.60**	88.40	77.20	84.62	↓1.45	54.50	↓**19.70**	8.00	↓2.00	6.80
R1b	92.40	92.00	86.40	86.80	86.00	84.40	61.60	85.20	90.80	88.80	77.60	84.73	↓1.35	67.80	↓6.40	8.00	↓2.00	7.60
R2	91.20	90.80	**87.60**	86.80	86.80	84.40	61.20	83.60	90.00	87.60	77.60	84.33	↓**1.74**	67.50	↓6.70	8.30	↓1.70	7.20
R3	92.80	91.60	86.80	87.20	87.20	84.40	57.20	85.20	90.80	89.20	78.00	84.58	↓1.49	69.60	↓4.60	7.90	↓**2.10**	7.20
R4	94.00	**93.20**	86.00	87.60	87.60	84.00	58.40	82.80	91.20	88.80	78.80	84.76	↓1.31	69.10	↓5.10	8.00	↓2.00	6.80
Key Hyperparameters																		
Rα	95.20	92.00	85.60	86.00	87.60	84.40	61.60	84.40	**91.60**	**90.00**	76.80	85.02	↓1.05	73.50	↓0.70	10.10	↑0.10	7.20
R$_\lambda$	95.60	91.60	86.40	86.80	88.00	81.60	53.60	73.20	88.40	84.00	59.60	80.80	↓**5.27**	64.90	↓**9.30**	3.90	↓**6.10**	6.80
R$_{\text{cov}}$	95.60	92.00	86.80	87.60	**88.40**	84.00	57.20	**86.80**	90.80	**90.00**	77.20	85.13	↓0.95	72.70	↓1.50	10.60	↑0.60	**9.20**
Training Strategies																		
R_{sft}	48.00	51.60	48.00	50.80	41.20	34.40	18.00	51.20	56.80	41.20	34.40	43.24	↓**42.83**	40.00	↓**34.20**	5.40	↓4.60	1.60
$R_{\text{sft+rl}}$	78.80	68.80	66.00	65.20	66.80	56.80	26.00	50.80	70.80	63.20	37.60	59.16	↓26.91	48.40	↓25.80	4.70	↓**5.30**	2.40

Core Components. We first incrementally add behavioral rewards and then introduce anchoring loss. R1 uses correctness reward, R1b adds discrete group filtering. Group filtering slightly improves AVG from 84.62% to 84.73%, and substantially increases $\overline{A}$ from 54.5% to 67.8%. R2 adds format and quality rewards, while R3 instead adds the language rewards. R3 achieves the highest $\overline{A}$ among the variants without anchoring (69.6%), showing the contribution of explicit language control. R4 combines all rewards and reaches 84.76% AVG. Finally, adding the anchoring loss yields Anchor, improving R4 by 1.31% in AVG, 5.10% in $\overline{A}$, and 2.00% in $\overline{M}$. Thus, behavioral rewards shape target output behavior, while representation anchoring boosts accuracy and language control.

Key Hyperparameters. We separately vary the anchoring lower bound α, weight λ, and coverage of Anchor. Lowering α from 0.7 to 0.5 causes only a modest drop of 1.05% in AVG and 0.70% in $\overline{A}$.

In contrast, increasing λ from 0.01 to 0.10 leads to substantial drops of 5.27%, 9.30%, and 6.10% in AVG, $\overline{A}$, and $\overline{M}$, indicating that overly strong anchoring can interfere with flexible reasoning. Extending anchoring from the latter half to the full reasoning segment reduces AVG by 0.95% and $\overline{A}$ by 1.50%, while increasing $\overline{M}$ by 0.60%. This indicates wider anchoring coverage shifts the trade-off between language control and multilingual reasoning, without accuracy gains.

Training Strategies. We compare SFT and SFT + RL strategies. The SFT variant drops 42.83% in AVG, with decreases in $\overline{A}$ and $\overline{M}$. Based on this SFT model, subsequent Anchor RL increases AVG from 43.24% to 59.16%, but still lower than Anchor. This implies SFT-model is unsuitable as a cold-start initialization, restricting its exploration space; we elaborate on this in the next section.

3.4 Does Anchor Work? Initialization and Recovery-Band Diagnosis

To examine Anchor's benefits across training starting points and their consistency with the intended representation-anchoring mechanism, we analyze from the three dimensions(Figure 3).

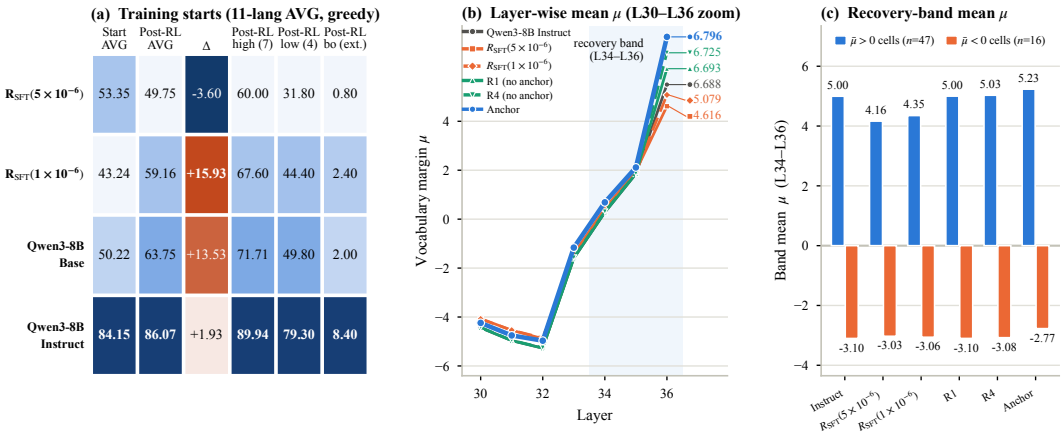

Figure 3: Initialization and recovery-band diagnosis. (a) Four initializations before and after full Anchor training on MGSM (Tibetan extension); colors are normalized within columns, and Δ is a within-row change. (b) Mean layer-wise vocabulary margin μ for six models on the 21 Anchor-eligible languages, zoomed to L30–L36; shading marks the recovery band. (c) Recovery-band means over cells with positive versus negative initial margins ($n = 47/16$).

Benefits across training starting points. Figure 3(a) evaluates Anchor across different RL starting points on MGSM. Outside cold-start setups, Anchor yields consistent accuracy gains. Starting from the strong Qwen3-8B-Instruct checkpoint, Anchor lifts accuracy by 1.93% to 86.07%, achieving the best overall performance. Starting from Qwen3-8B Base delivers a larger 13.53% gain but a lower final accuracy of 63.75%, showing Anchor's gain magnitude depends on the starting checkpoint capability. For cold-start initialization, consistent with Son et al. (2026), the initial checkpoint is substantially weaker than standard backbones. At the lower learning rate, Anchor gains 15.93% to 59.16%, though still below non-cold-start runs; the higher learning rate incurs a 3.60% performance drop. This indicates Anchor performs best when the starting checkpoint preserves sufficient exploration space, with larger absolute performance for stronger base models.

Recovery-band representation analysis. Figure 3(b) shows the average vocabulary margin μ in the recovery band across models, computed over the 21 anchored languages and three recovery-band layers (Appendix B), yielding 63 representation units. Except for Anchor, the other five models exhibit similar μ distributions, without a consistent positive shift, suggesting that conventional RL or behavioral language rewards alone do not substantially alter the language-routing in the recovery band. Compared with R4(with all rewards but no anchoring loss), Anchor produces a consistent positive shift in μ together with improved accuracy. This indicates that the representation-level anchoring mechanism operates as intended. Figure 3(c) further examines the effect across different initial margin μ state. Among the 63 representation units, 47 units hold positive μ (existing language advantage), and 16 hold negative μ (weak or absent language advantage). Relative to Qwen3-8B-Instruct, for positive-μ units, Anchor yields μ 5.23, while cold-start models drop to 4.16 and 4.35;

for negative-μ units, Anchor lifts μ to -2.77, whereas all non-anchored models remain near -3.10. Anchor achieves the expected anchoring effect by restoring the negative μ language to the lower anchoring limit while retaining the existing language routing of the positive μ language.

3.5 Why Does Anchor Work? Language Behavior and Anchoring Dynamics

We further track how training affects the behavioral and representation levels (Table 3 and Figure 4).

Table 3: Language behavior on MGSM. A/M use fixed all-response denominators. $p_o = 1 - p_g - p_{en}$ is the residual share of languages other than the question language and English.

Model	A (alignment)	M (mixing)	p_g	p_{en}	p_o	Empty-public rate
Qwen3-8B-Instruct	49.0	8.0	5.7	87.7	6.6	43.5
R1 (correctness only)	54.5	8.0	5.7	87.7	6.6	36.9
R4 (all rewards, no anchoring)	69.1	8.0	7.1	86.5	6.4	20.9
Anchor	**74.2**	**10.0**	**7.5**	85.9	6.6	**15.2**

Language behavior dynamics. Table 3 summarizes language behavior across training configurations. Relative to the Qwen3-8B-Instruct, R1 (correctness reward) lifts alignment A from 49.0% to 54.5%, and R4 (full rewards) further raises it to 69.1%. With representation anchoring, Anchor pushes $\overline{A}$ to 74.2% and $\overline{M}$ to 10.0%. On reasoning-segment language distribution, p_g rises by 1.8% and p_{en} drops by 1.8% relative to the backbone, while p_o remains at 6.6%, denoting anchoring recovers question-language routing, boosting public-segment alignment and question-language participation in reasoning. Further decomposition shows remaining languages distribute across Spanish, French, and others with no consistent directional preference (Appendix C.4). Figure 4(d) visualizes the evolution of $\overline{A}$ and $\overline{M}$ during training. Both metrics grow stably with training iterations, supporting the results in Table 3, confirming that Anchor steadily boosts public-segment alignment and promotes question-language usage in reasoning, while maintaining flexible multilingual reasoning.

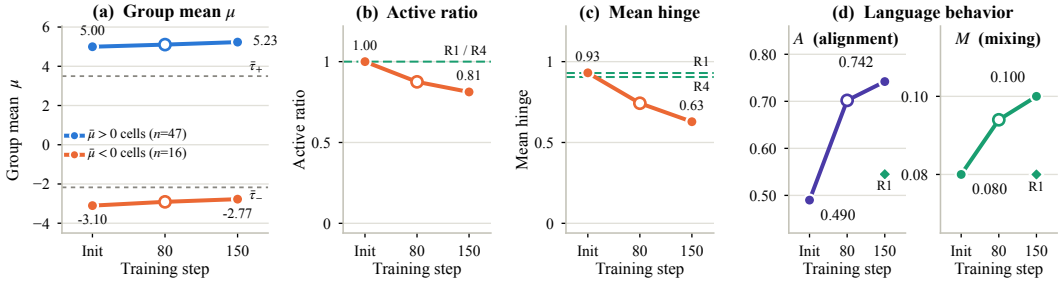

Figure 4: Anchor dynamics at initialization, steps 80 and 150. (a) Vertical axis: group mean vocabulary margin μ, grouped by initial margin sign; dashed lines mark group-mean bounds at $\alpha = 0.7$. (b) Active ratio: fraction of 16 negative-initial-margin cells with $\mu < \tau$. (c) Mean hinge: average $\max(0, \tau - \mu)$ over these 16 cells, with zeros for inactive cells. Horizontal references in (b,c) show final R1/R4 values. (d) Public alignment A and reasoning mixing M; diamonds mark R1.

Anchoring dynamics. Figure 4 tracks the evolution of the 63 recovery-band representation units introduced in Section 3.4. Figure 4(a) shows mean μ rises steadily for units with and without initial routing advantage: positive-advantage units grow from 5.00 to 5.23, while disadvantaged units improve from -3.10 to -2.77, matching the final patterns in Figure 3(c). This confirms Anchor's representation gains stem directly from the anchoring objective. Figures 4(b) and (c) focus on initially 16 disadvantaged units. We define the *active ratio* as the fraction of units satisfying $\mu < \tau$ (below the anchoring lower bound). This ratio falls from 1.00 to 0.81 at step 150. The *mean hinge*, defined as the average $\max(0, \tau - \mu)$ for these units and measuring distance to the bound, drops from 0.93 to 0.63. For non-anchored R1 and R4, these two metrics stay nearly fixed. These dynamics confirm that the anchoring loss works as designed: it gradually optimizes recovery-band representations and restores routing for initially disadvantaged languages.

3.6 Cross-Benchmark Generalization

We further evaluate Anchor on knowledge-oriented benchmarks: MMLU-ProX-Lite and Global-MMLU-Lite. (Figure 5). Anchor achieves the highest accuracy on both benchmarks: 70.31% on Global-MMLU-Lite and 55.28% on MMLU-ProX-Lite. On Global-MMLU-Lite, Anchor outperforms Qwen3-8B-Instruct (70.23%) by 0.08%, with slim gains over larger baselines including Gemma 3 IT 12B (69.50%), indicating saturated benchmark performance. On MMLU-ProX-Lite, it surpasses Llama-3.3-70B by 2.38% and Qwen3-8B-Instruct by 3.27%, demonstrating strong cross-benchmark generalization at small parameter scales.

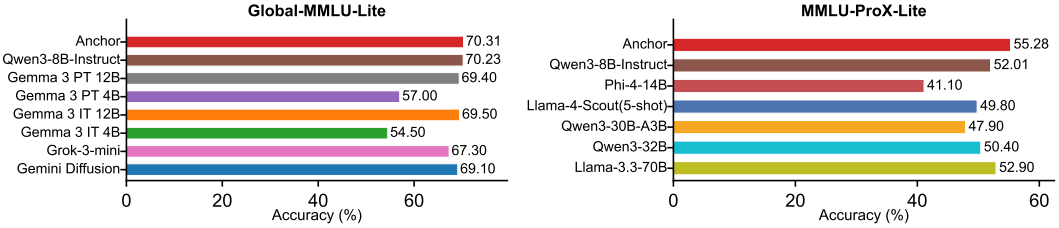

Figure 5: Cross-Benchmark accuracy, full results in Appendix D ($\overline{A}$, $\overline{M}$ and per-language results).

4 Related Work

Multilingual Reasoning and Output Language Control. MGSM (Cobbe et al., 2021; Shi et al., 2022), Global-MMLU-Lite (Singh et al., 2024), and MMLU-ProX-Lite (Xuan et al., 2025) consistently show that reasoning performance degrades with language resource level. Failing to respond in the question language is also a distinct multilingual failure (Mukherjee et al., 2026). Reasoning language presents a trade-off between accuracy and language fidelity. English reasoning can improve accuracy (Shi et al., 2022; Weihua et al., 2026), whereas enforcing the question language can hurt low-resource reasoning (Yong et al., 2025; Saji et al., 2026), with performance in some language–script settings dropping to near-collapse levels under script-constrained decoding(Wang et al., 2025). Meanwhile, English reasoning can introduce errors through mistranslation of the problem statement (Saji et al., 2026). Language mixing increases with task difficulty and closely tracks hidden-layer language composition (Wang et al., 2025), while language-mixed CoT distillation outperforms both monolingual extremes in most settings (Son et al., 2026). Existing approaches address this trade-off through prompting and decoding interventions (Yong et al., 2025), language-consistency rewards (Guo et al., 2025; Park et al., 2026), language-mixed CoT distillation (Son et al., 2026), and external-judge-based routing (Weihua et al., 2026).

RL for Reasoning and Language-Related Representations. GRPO (Shao et al., 2024) uses group-relative advantages instead of a value network, while DAPO (Yu et al., 2026) improves long-CoT RL stability. In multilingual RL, DeepSeek-R1 introduces language consistency rewards (Guo et al., 2025), (Park et al., 2026) show that correctness driven RL shifts reasoning toward the dominant pre-training language, revealing a persistent accuracy–language-fidelity trade-off. IRIS (Gupta et al., 2026) improves low-resource reasoning with staged training and composite rewards, but still relies on English data to sustain the reward signal. Language behavior is reflected in representations, (nostalgebraist, 2020; Wendler et al., 2024; Tang et al., 2024) show that middle layers tend toward shared representations, while late layers re-specialize to individual languages . Hidden-layer language composition also closely tracks reasoning language (Wang et al., 2025), indicating that multilingual behavior has both behavioral and representational dimensions.

5 Conclusion

We investigate the balance between correctness and language in multilingual long-CoT reasoning, which manifests at both behavioral and representational levels. To resolve this issue, we propose Anchor, regulating output language behavior via multidimensional rewards and preserves recovery-band vocabulary routing through representation anchoring loss. Experiments across multiple benchmarks verify that our method effectively addresses the accuracy–language trade-off.

LIMITATIONS

Our accuracy results cover MGSM (with the Tibetan extension) and two knowledge benchmarks, Global-MMLU-Lite and MMLU-ProX-Lite. The anchoring gain transfers to MMLU-ProX-Lite, but no significant accuracy change is detected on Global-MMLU-Lite. The gains of the method on knowledge tasks are therefore not universal.

The restricted syntax of the canonical answer segment is a single number or a single option letter. This enables automatic grading. It does not cover tasks whose answers are free text, proofs, or code. Extending to such settings needs a new answer syntax and a new grading protocol. We do not verify this in this paper.

The vocabulary margin used by the anchoring loss is a proxy signal. It compares the log probability mass of two token sets. It is not a full language classifier. The initial margin is measured by an input-side probe and transferred to a generation-side per-token constraint. The probe positions, in the second half of the question, and the constraint positions, in the second half of the reasoning segment, do not coincide. The effect may depend on token set size, dual-script languages, and position distribution.

Under a negative initial margin, the unified formula shrinks the margin toward zero. This is a correction, not a pure preservation. Results should be reported by language-by-layer groups. They should not be read as complete language ability or as an independent causal conclusion.

The language signals depend on the coverage of fastText. Six of the 29 Latin-script training languages—ha, ig, ny, sn, wo, and zu—do not receive language rewards. The language behavior metrics A and M also build on fastText probabilities and on chunk cleaning rules. For Tibetan and Swahili, the effective denominator of the language-behavior metrics is limited. We report the related conclusions under macro-average aggregation.

The Tibetan data are a small human-translated extension. Appendix A.2 describes its size, translation, and validation protocol.

We use a single 8B base model, a single epoch, and a single seed per configuration. The SFT initialization comparison and the recovery-band diagnosis are both correlational evidence.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation]. We have not used generative AI tools for [generate synthetic data sets; design or provide feedback on research methodology or experiments; implement methods; clean and reformat dataset; interpret results], and [help develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in the writing of proofs; propose or refine hypotheses; support qualitative and thematic data analysis] are not applicable to this work. Additionally, we used generative AI tools for [edit a research paper to improve readability]. We have reviewed all AI-assisted work. [Multiple authors reviewed and post-edited the manuscript over multiple rounds, with particular attention to the accuracy of the translation, the clarity of the language, and the consistency of the content, to ensure that the final manuscript accurately reflects the authors' intended content and contributions.] We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations, 2023.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Navya Gupta, Rishitej Reddy Vyalla, Avinash Anand, Chhavi Kirtani, Erik Cambria, Zhengchen Zhang, Zhengkui Wang, Timothy Liu, Aik Beng Ng, Simon See, et al. Iris: Interleaved reinforcement with incremental staged curriculum for cross-lingual mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22216–22248, 2026.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016a.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016b.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International conference on machine learning*, pp. 22631–22648. PMLR, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Anjishnu Mukherjee, Chutong Meng, and Antonios Anastasopoulos. Lost in the tower of babel: The adverse effects of incidental multilingualism in llms. *arXiv preprint arXiv:2605.01224*, 2026.

nostalgebraist. Interpreting GPT: the logit lens, August 2020. LessWrong blog post. Available at `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, retrieved September 2026.

Cheonbok Park, Jeonghoon Kim, Joosung Lee, Sanghwan Bae, Jaegul Choo, and Kang Min Yoo. When verifiable rewards switch the language: Cross-lingual collapse in chain-of-thought. 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Alan Saji, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. The reasoning lingua franca: A double-edged sword for multilingual ai. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 329–344, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

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, 2022.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024. URL https://arxiv.org/abs/2412.03304.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Srikant Panda, Minhyuk Kim, Dasol Choi, Kyong-Ha Lee, Youngjae Yu, et al. Pushing on multilingual reasoning models with language-mixed chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 79387–79407, 2026.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schütze. Language mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2637–2665, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zheng Weihua, Xin Huang, Zhengyuan Liu, Tarun Kumar Vangani, Bowei Zou, Xiyan Tao, Yuhao Wu, AiTi Aw, Nancy F Chen, and Roy Ka-Wei Lee. Adamcot: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33863–33871, 2026.

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

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*, 2025.

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

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

A Training Data, Initializations, and Evaluation Protocol

This appendix collects the experimental-protocol details: the construction of the 50-language training corpus, the Tibetan translation and validation protocol, the distilled SFT initializations, the applicability of the fastText language signals, and the evaluation settings and prompt templates.

A.1 Training-Data Construction

Multilingual parallel MGSM8KInstruct (numeric questions). We merge parallel MGSM8KInstruct instruction data in ten languages (bn, de, en, es, fr, ja, ru, sw, th, zh) with our Tibetan subset, deduplicate at the semantic-group level by hashes of normalized question text, remove questions whose reference answers match none of the five supported numeric forms, and stratify by per-language availability to sample complete semantic groups, yielding 11 languages including bo.

Global-MMLU (multiple-choice questions). After removing the Global-MMLU-Lite evaluation subset, integer linear programming balances subject category (STEM, humanities, social sciences, medicine, business, other) and cultural-sensitivity ratio over the English-version candidate pool; all 42 language versions of each selected group are retained (consolidated in Table 4).

MMLU-ProX (multiple-choice questions). We remove the MMLU-ProX-Lite evaluation subset, stratify the remainder by 14 subject categories, and retain all 29 language versions of each selected group.

Tibetan subset. The subset contains about 2.3k questions and a parallel Tibetan test set built under the MGSM protocol; Appendix A.2 documents the translation and validation protocol.

Splitting and leakage checks. We sort semantic groups by a stable hash and split complete groups within each source-by-question-type stratum into 14,472 training prompts (5,168 groups) and 771 dev prompts (286 groups), approximately 95%/5%. The dev set is used for validation, model selection, and early stopping. Question identifiers and question-text hashes are checked against the test items of MGSM with the Tibetan extension, Global-MMLU-Lite, and MMLU-ProX-Lite; any nonzero overlap aborts data production.

Language inventory. Table 4 consolidates the language sets. The 50 training languages comprise 21 non-Latin-script languages—all Anchor-eligible and fastText-supported— and 29 Latin-script languages, 23 of which are fastText-supported. The derived sets follow from this partition: the layer-wise probe covers 26 languages (the 21 plus the Latin-script evaluation languages en, es, fr, de, sw); the MGSM evaluation covers 12 languages, of which seven are Anchor-eligible (zh, ja, ru, th, bn, te, bo); Anchoring coverage on the Lite benchmarks is 6 of 23 Global-MMLU-Lite languages (ar, bn, hi, ja, ko, zh) and 14 of 29 MMLU-ProX-Lite languages (ar, bn, hi, ja, ko, mr, ne, ru, sr, te, th, uk, ur, zh).

Table 4: Language inventory. The 50 training languages partition by script; Anchoring and fastText applicability follow from this partition.

Group	#	Languages	Anchor	fastText
Non-Latin	21	zh, ja, ko, ru, uk, ky, sr, el, th, bo, he, hi, mr, ne, bn, si, te, am, ar, fa, ur	yes	yes
Latin, supported	23	en, es, fr, de, sw, cs, fil, id, it, lt, mg, ms, nl, pl, pt, ro, so, sv, tr, vi, yo, af, hu	–	yes
Latin, unsupported	6	ha, ig, ny, sn, wo, zu	–	–

A.2 Tibetan Translation and Validation Protocol

Translation guidelines. The Tibetan training subset and the parallel Tibetan MGSM test questions were translated from the original English source data to preserve source meaning without unsupported omissions or additions. Annotators were required to translate from the English originals rather than from other language versions, to keep proper nouns such as person and place names in English, and to follow the annotation format of the original dataset. Translators used standard written Tibetan for domain-specific terminology, including religious expressions, and checked grammar and fluency. Ambiguous or culturally specific wording was cross-checked and resolved by annotator consensus.

Consistency check. In a consistency check over 50 sampled translated items, 48 of the 50 (96%) were consistent.

A.3 Distillation Details

Distilled SFT initializations. We train two distilled SFT models on the same self-distilled corpus using learning rates of 5×10^{-6} and 1×10^{-6}, denoted by $R_{\mathrm{SFT}}(5 \times 10^{-6})$ and $R_{\mathrm{SFT}}(1 \times 10^{-6})$, respectively. These checkpoints are evaluated as SFT baselines and are also used as alternative initializations for Anchor RL. The SFT corpus is disjoint from the RL corpus used in this paper. It covers 34 languages with 13,490 training samples and 1,349 validation samples. Its questions come from parallel MGSM8KInstruct math problems and from Flan-CoT(Longpre et al., 2023) multiple-choice and free-response questions that have been translated into multiple languages.

Trajectory construction. For each question, the initialization model itself (Qwen3-8B-Instruct) generates the reasoning trajectory. We keep only its reasoning segment as the internal reasoning chain, then append the original target-language solution provided with the dataset. For math and multiple-choice questions we further append a single canonical answer marker "####" on the last line; free-response questions omit this marker.

Filtering and leakage control. A trajectory is retained only if it passes thresholds on format, length (96–4,096 tokens for the reasoning segment and at most 8,192 tokens for the whole sample), repetition, and answer correctness, with the correctness threshold relaxed for Tibetan. Samples are partitioned by semantic group, and any question overlapping the three evaluation sets is removed.

No question-level overlap with the RL corpus. The SFT and Anchor-RL corpora do not overlap at the question level. RL math questions are sampled from the same candidate pool as SFT after removing, by question hash, every item already consumed by SFT; the RL multiple-choice sources (Global-MMLU and MMLU-ProX) are not used by SFT at all.

A.4 Applicability of the fastText Language Signals

Language discrimination uses label probabilities from the fastText language-identification model lid.176. We first split text coarsely at punctuation and line breaks. We remove numbers, formulas, URLs, code, option labels, and structural markers. We retain only chunks with at least 20 alphabetic characters and weight all retained chunks equally. A language signal is considered applicable when the question language is in the model's supported label set. This is only a proxy for label availability. It does not mean that identification accuracy has been independently calibrated.

Among the 50 training languages, 44 are supported. The six exceptions—ha, ig, ny, sn, wo, and zu, all Latin-script—are outside the supported set; for these languages, public-segment alignment and reasoning-segment mixing are set to zero and produce no language reward. Correctness, format validity, repetition, and overlength are computed in the same way as for other languages. English is supported, so public-segment alignment can be computed. We do not compute an English–English mixing reward.

A.5 Evaluation Settings

MGSM with the Tibetan extension contains 250 questions per language. We evaluate models through a service deployed with vLLM(Kwon et al., 2023). The evaluation instruction uses different wording from training but preserves the same output contract. The template prescribes no reasoning

language; the public-segment language requirement is the only language guidance. We use three decoding settings: greedy decoding ($T = 0$), temperature sampling ($T = 0.6$), and temperature 0.6 with repetition penalty 1.05. The main tables use greedy decoding. Numeric answers are compared by exact value under the restricted syntax. Supported forms include integers, decimals, fractions, mixed numbers, and percentages. Valid thousands separators are normalized. Multiple-choice answers are compared by exact option letter. Appendix C.1 reports per-language results at steps 80 and 150. The main text only reports the end-of-training checkpoint at step 150. Published external baselines are not evaluated on Tibetan.

For Global-MMLU-Lite, we evaluate 9,199 questions in 23 languages: 399 in my and 400 in each other language. For MMLU-ProX-Lite, we evaluate 588 questions in each of 29 languages, for 17,052 questions in total. We evaluate five core rows: Qwen3-8B-Instruct, $R_{\mathrm{SFT}}(5 \times 10^{-6})$, R1, R4, and Anchor. Each is evaluated once with greedy decoding and once with temperature 0.6 plus repetition penalty 1.05. Other settings match MGSM. The instruction style and output contract are the same, and grading uses the final `#### <option letter>` marker.

The main RL experiment uses seed 42, with checkpoints saved at steps 40, 80, 120, and 150.

A.6 Prompt Templates and Output-Contract Examples

Training and evaluation use two instructions with different wording but the same contract. The reasoning segment allows any useful language or mixture. The public segment uses the question language. The last line contains the canonical answer after `####`.

Training template. This template follows the question. Numeric questions use `#### <number>`; multiple-choice questions use `#### <option label>`. Multiple-choice questions use `Question:` and `Options:`, with one labeled option per line.

```
{question}

Inside <think></think>, use any language or mixture
of languages that helps solve the problem.
After </think>, explain the solution in the same
language as the question. On the final line, output
#### <number> / #### <option label>.
```

MGSM evaluation template. Here `{lang}` is the English name of the question language. This template is used for all results in Table 1 and Appendix C.1.

```
/think

Solve the following math problem step by step.

- Your reasoning process (inside <think></think>)
  may be written in any language or a mix of
  languages. Choose whichever language(s) help
  you solve the problem best.
- Your final answer (after </think>) MUST be written in {lang}.
- The problem is presented in {lang}; do not
  translate the problem text.

Problem: {question}

Output the final numerical answer in the format: #### <number>
```

Multiple-choice evaluation template. Global-MMLU-Lite and MMLU-ProX-Lite use this variant. Options are formatted as in the training data and labeled A through J on separate lines.

```
/think
```

```
Answer the following multiple choice question.

- Your reasoning process (inside <think></think>)
  may be written in any language or a mix of
  languages. Choose whichever language(s) help
  you solve the problem best.
- Your final answer (after </think>) MUST be written in {lang}.
- The problem is presented in {lang}; do not
  translate the problem text.

Question: {question}

Options:
A. {option_a}
B. {...}

Output the final answer in the format: #### <option letter>
```

Response-parsing contract. A response must contain exactly one ordered pair of `<think>` and `</think>` markers. A nonempty public segment must follow. The last line must contain the only `####` answer marker. The format is invalid ($f = 0$) when the think markers are missing, duplicated, or out of order; when the public segment is empty; when the answer marker is not unique or cannot be parsed; or when text follows the canonical answer. In the last case, correctness is also set to zero because the canonical answer must end the response.

B Representation-Level Probe and Anchoring Bounds

The layer-wise probe covers 26 languages with 100 examples per language, for a total of 2,600 probe examples. The probe set is constructed in two batches. The first batch covers an 11-language parallel mathematics pool comprising six non-Latin-script languages (zh, ja, ru, th, bn, bo) and five Latin-script control languages (en, de, es, fr, sw). These examples are drawn from the multilingual parallel mathematics-question pool that shares its upstream source with the SFT data. The second batch covers the remaining 15 non-Latin-script languages in the Anchor-RL training distribution (am, ar, el, fa, he, hi, ko, ky, mr, ne, si, sr, te, uk, ur). For each language, 100 examples are sampled from the Anchor-RL data using language-stratified sampling with a fixed random seed (seed $= 42$); their upstream questions originate from Global-MMLU (Singh et al., 2024) and MMLU-ProX (Xuan et al., 2025). Both batches use structurally identical Alpaca-style templates, and each probe record contains only the question prompt, with no generated response. The representation analyses use the 21 non-Latin-script languages, whereas the five Latin-script languages are included only as script-level controls.

Table 5 reports the recovery-band vocabulary margins $\bar{\mu}_{l,g}$ probed on the initial Qwen3-8B-Instruct over the 21 Anchor-eligible languages; the Anchoring bound is $\tau_{l,g} = 0.7\bar{\mu}_{l,g}$, applied uniformly to every anchor-eligible language and layer. The Latin- versus non-Latin-script comparison in the main text draws on the complete 26-language probe; absence from this table means only that no separable question-language token set, and hence no Anchoring loss, exists for that language, not that the language has no layer-wise signal.

Because $\tau - \bar{\mu} = -(1-\alpha)\bar{\mu}$, a positive initial margin yields a bound below the initial value (a preserving bound), whereas a negative initial margin yields a bound above it (a corrective bound). For Serbian at L_{34}, $\bar{\mu} = -5.88$ and $\tau = -4.12$: a margin that stays at -5.88 incurs a hinge of 1.76 and a gradient toward zero, which vanishes at -4.12; the loss never forces a positive margin. For sr, whose margins are negative in all three layers, Anchoring mainly prevents further degradation relative to the initial model, and positive improvement relies on the behavioral rewards. Consequently, the effect of α depends on the sign—larger α tightens positive-margin bounds but weakens negative-margin ones—so the α ablation must be interpreted separately by sign group. Some scripts have small question-language token sets (bn: about 47 tokens; bo: about 44), so absolute margins are interpreted only as relative trends within the same language, token set, and layer.

Table 5: Recovery-band vocabulary margins $\bar{\mu}_{l,g}$ in the initial model. The Anchoring bound is $\tau_{l,g} = 0.7\bar{\mu}_{l,g}$.

Language	L_{34}	L_{35}	L_{36}	**Language**	L_{34}	L_{35}	L_{36}
zh	8.11	5.71	9.42	hi	3.85	4.92	9.09
ja	4.25	4.28	8.38	mr	2.74	3.38	7.11
ko	-0.66	-0.52	3.09	ne	3.12	3.96	6.76
ru	2.65	4.21	11.85	bn	0.12	3.28	8.56
uk	1.04	3.17	9.30	si	0.71	1.50	4.34
ky	0.40	1.82	8.30	te	-5.20	-3.59	2.07
sr	-5.88	-5.18	-2.91	am	-3.35	-2.90	1.56
el	4.25	7.20	13.07	ar	-4.40	-1.63	4.99
th	-2.18	4.51	11.50	fa	-0.43	2.64	7.38
bo	4.22	3.79	7.83	ur	-0.86	1.65	5.88
he	-6.64	-3.31	2.89				

Table 6: Per-language accuracy (%) on MGSM with the Tibetan extension. AVG excludes bo, which is reported separately. Internal models receive no reasoning-language guidance. The three decoding settings are greedy ($T = 0$), temperature sampling ($T = 0.6$), and temperature 0.6 with repetition penalty 1.05. Table 1 uses only the step-150 greedy results.

Model	**Setting**	**en**	**es**	**fr**	**de**	**zh**	**ja**	**sw**	**bn**	**ru**	**th**	**te**	**bo**	**AVG**
Qwen3-8B-Instruct	T=0	92.40	90.80	85.20	87.60	86.40	83.60	57.60	85.60	90.00	88.00	78.40	7.20	84.15
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	T=0	68.40	68.00	64.40	63.20	66.40	50.00	23.60	44.40	70.40	50.80	17.20	1.20	53.35
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	T=0.6	75.20	67.20	65.60	70.40	68.40	52.80	23.20	43.20	70.80	46.80	23.60	2.00	55.20
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	rep=1.05	79.60	68.40	66.00	68.80	66.80	54.00	28.80	49.20	72.80	53.20	22.80	1.20	57.31
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	T=0	48.00	51.60	48.00	50.80	41.20	34.40	18.00	51.20	56.80	41.20	34.40	1.60	43.24
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	T=0.6	42.40	58.00	46.80	56.80	48.00	40.40	16.00	49.20	56.80	45.20	34.40	0.80	44.91
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	rep=1.05	40.80	53.20	46.80	54.00	43.20	38.00	24.00	47.60	59.20	41.20	37.20	1.20	44.11
Anchor ($R_{\mathrm{SFT}}(5 \times 10^{-6})$ init., step150)	T=0	68.00	58.40	59.20	59.20	61.20	47.20	23.20	36.00	66.80	48.00	20.00	0.80	49.75
Anchor ($R_{\mathrm{SFT}}(5 \times 10^{-6})$ init., step150)	T=0.6	71.20	66.80	60.80	63.60	61.60	50.40	22.80	48.40	66.80	54.80	22.40	0.40	53.60
Anchor ($R_{\mathrm{SFT}}(5 \times 10^{-6})$ init., step150)	rep=1.05	76.40	70.80	69.20	68.00	65.20	55.60	25.20	46.80	69.60	55.60	30.40	1.20	57.53
Anchor ($R_{\mathrm{SFT}}(5 \times 10^{-6})$ init., step80)	T=0	70.00	59.20	62.80	61.20	61.20	48.40	20.40	37.60	67.20	50.40	20.40	1.20	50.80
Anchor ($R_{\mathrm{SFT}}(5 \times 10^{-6})$ init., step80)	T=0.6	78.00	66.00	62.00	66.00	60.80	52.40	23.60	38.40	66.40	55.60	23.20	1.20	53.85
Anchor ($R_{\mathrm{SFT}}(5 \times 10^{-6})$ init., step80)	rep=1.05	79.20	68.00	64.80	67.60	68.00	56.80	28.80	50.40	66.80	64.80	26.00	1.60	58.29
Anchor ($R_{\mathrm{SFT}}(1 \times 10^{-6})$ init., step150)	T=0	78.80	68.80	66.00	65.20	66.80	56.80	26.00	50.80	70.80	63.20	37.60	2.40	59.16
Anchor ($R_{\mathrm{SFT}}(1 \times 10^{-6})$ init., step150)	T=0.6	82.40	72.80	68.80	70.40	72.40	61.60	26.00	56.40	75.20	63.60	34.80	0.80	62.22
Anchor ($R_{\mathrm{SFT}}(1 \times 10^{-6})$ init., step150)	rep=1.05	82.80	76.00	70.00	74.00	72.40	67.20	34.00	57.60	77.60	72.40	40.00	1.60	65.82
Anchor ($R_{\mathrm{SFT}}(1 \times 10^{-6})$ init., step80)	T=0	76.40	68.40	65.20	66.80	66.40	52.40	28.00	51.60	72.40	65.60	38.40	0.80	59.24
Anchor ($R_{\mathrm{SFT}}(1 \times 10^{-6})$ init., step80)	T=0.6	80.40	78.00	70.80	72.80	69.20	63.20	24.00	54.80	76.00	65.20	36.80	1.60	62.84
Anchor ($R_{\mathrm{SFT}}(1 \times 10^{-6})$ init., step80)	rep=1.05	79.60	77.20	74.00	76.00	70.40	67.60	31.60	61.60	80.80	71.60	45.60	4.40	66.91
Anchor (Instruct init., step 80)	T=0	94.80	92.40	86.80	88.00	88.40	84.40	58.80	87.20	92.40	89.20	77.20	6.40	85.42
Anchor (Instruct init., step 80)	T=0.6	93.20	92.40	86.80	88.40	89.20	86.40	60.40	83.60	91.60	87.20	78.00	8.00	85.20
Anchor (Instruct init., step 80)	rep=1.05	95.20	92.40	86.40	87.20	89.20	85.20	60.00	86.40	93.20	90.40	79.20	12.80	85.89
Anchor (Instruct init., step 150)	T=0	96.00	93.20	86.40	88.80	88.40	86.00	62.00	85.20	90.80	89.20	80.80	8.40	86.07
Anchor (Instruct init., step 150)	T=0.6	95.20	92.40	88.40	88.00	88.80	84.40	60.00	85.20	91.60	88.80	79.20	8.80	85.64
Anchor (Instruct init., step 150)	rep=1.05	93.60	93.20	88.00	90.40	88.40	85.20	64.40	87.60	91.60	88.00	79.60	8.40	86.36
Qwen3-8B-Base (initialization itself)	T=0	84.00	66.00	61.20	66.80	66.00	36.40	12.40	27.60	54.80	69.60	7.60	1.20	50.22
Qwen3-8B-Base (initialization itself)	rep=1.05	60.00	54.00	58.40	59.60	56.00	45.60	14.00	26.40	57.20	54.80	16.80	3.60	45.71
Anchor (Base init., step 80)	T=0	83.60	64.40	56.40	56.00	78.00	62.40	17.60	45.20	66.80	73.20	50.00	1.60	59.42
Anchor (Base init., step 80)	T=0.6	73.60	64.40	54.80	62.00	73.20	56.40	17.20	46.00	62.80	65.20	38.40	2.40	55.82
Anchor (Base init., step 80)	rep=1.05	71.20	58.40	56.40	62.40	71.60	55.20	19.60	41.20	64.40	63.20	40.00	2.80	54.87
Anchor (Base init., step 150)	T=0	82.00	72.80	60.40	62.00	82.80	69.20	17.60	50.00	72.80	75.60	56.00	2.00	63.75
Anchor (Base init., step 150)	T=0.6	83.60	71.20	58.00	66.40	78.80	64.40	20.00	51.20	73.60	69.60	46.00	4.80	62.07
Anchor (Base init., step 150)	rep=1.05	78.40	70.80	64.40	61.60	73.60	65.60	17.20	48.80	74.40	66.00	46.00	3.20	60.62
s1-7B (English-reasoning baseline)†	Published	92.40	88.80	83.20	87.60	86.00	82.40	24.00	72.00	88.00	81.60	36.80	–	74.80
s1-7B (budget forcing, truncation)†	Published	88.80	86.00	82.00	82.80	83.60	78.80	21.60	65.20	86.40	80.00	38.80	–	72.18
s1-7B (budget forcing, extrapolation)†	Published	90.40	83.60	84.40	84.00	82.80	74.80	19.20	70.80	84.40	78.40	36.40	–	71.75
Qwen2.5-7B-Instruct †	Zero-shot	78.00	72.80	66.40	69.20	79.60	67.20	13.60	59.20	71.20	68.80	33.20	–	61.75

C SUPPLEMENTARY MGSM RESULTS

This appendix collects the supplementary MGSM results: per-language accuracy under the three decoding settings, full ablation results with paired significance tests, per-language language-behavior statistics, and the decomposition of the residual language share p_o.

C.1 Full Per-Language Results under Three Decoding Settings

The external models, s1-7B and Qwen2.5-7B-Instruct, were not evaluated on Tibetan, and their averages cover 11 languages. Their values come from (Yong et al., 2025). Qwen2.5-7B-Instruct and the two s1 budget-forcing rows use the exact per-language results in Table 5 of that paper. The s1 English-reasoning baseline uses the exact per-language Baseline row for s1.1-7B in its Table 11. The original paper reports an 11-language average of 74.8 for this row.

The best Tibetan value is 12.80 from Anchor with the Instruct initialization at step 80 under rep=1.05; this Tibetan extension is excluded from AVG. The Qwen3-8B-Base initialization and the corresponding Anchor rows match the base-initialization comparison in Figure 3(a).

C.2 Full Ablation Results and Significance Tests

Table 7: Per-language MGSM accuracy (%) for eight ablation configurations at the end-of-training checkpoint. AVG is the mean over the original 11 languages; bo is reported separately. Each configuration is evaluated with three decoding settings, with no reasoning language prescribed. R1b differs from R1 only in the dynamic-sampling filter. R_λ uses $\lambda = 0.10$. $\mathrm{R}_{\mathrm{cov}}$ uses reasoning-segment coverage 1.0.

Configuration	Setting	en	es	fr	de	zh	ja	sw	bn	ru	th	te	bo	AVG
R1	T=0	93.60	92.00	85.20	88.80	86.40	84.40	58.40	84.80	91.60	88.40	77.20	6.80	84.62
R1	T=0.6	93.20	91.20	87.20	89.60	88.00	86.40	57.20	85.60	91.20	89.20	77.20	8.00	85.09
R1	rep=1.05	92.80	90.40	86.80	88.80	87.20	84.80	58.40	86.40	92.00	89.60	77.20	8.00	84.95
R1b	T=0	92.40	92.00	86.40	86.80	86.00	84.40	61.60	85.20	90.80	88.80	77.60	7.60	84.73
R1b	T=0.6	94.00	90.80	87.20	86.40	88.00	84.40	60.00	83.60	92.80	88.80	76.80	9.60	84.80
R1b	rep=1.05	93.60	92.00	86.00	88.40	88.40	86.40	59.60	85.20	91.20	89.60	78.40	7.20	85.35
R2	T=0	91.20	90.80	87.60	86.80	86.80	84.40	61.20	83.60	90.00	87.60	77.60	7.20	84.33
R2	T=0.6	93.20	90.80	87.20	87.60	87.20	85.20	58.80	82.80	88.80	87.60	77.20	10.80	84.22
R2	rep=1.05	94.00	92.80	88.40	87.20	88.00	84.40	61.60	88.00	92.40	90.40	78.40	10.40	85.96
R3	T=0	92.80	91.60	86.80	87.20	87.20	84.40	57.20	85.20	90.80	89.20	78.00	7.20	84.58
R3	T=0.6	93.60	91.60	88.00	87.20	87.20	84.40	58.40	84.40	91.60	87.60	78.40	7.20	84.76
R3	rep=1.05	94.40	92.80	87.20	88.40	88.00	84.00	61.60	87.20	92.00	88.00	79.60	9.60	85.75
R4	T=0	94.00	93.20	86.00	87.60	87.60	84.00	58.40	82.80	91.20	88.80	78.80	6.80	84.76
R4	T=0.6	93.20	90.40	86.00	88.00	87.60	86.80	54.40	83.60	91.60	89.20	76.00	6.00	84.25
R4	rep=1.05	92.00	92.40	87.60	90.00	88.40	84.80	61.60	86.40	92.00	87.20	78.00	11.60	85.49
Rα	T=0	95.20	92.00	85.60	86.00	87.60	84.40	61.60	84.40	91.60	90.00	76.80	7.20	85.02
Rα	T=0.6	95.20	91.20	86.40	88.40	88.00	86.00	59.60	84.00	92.80	88.80	78.00	8.00	85.31
Rα	rep=1.05	94.40	92.40	88.80	88.00	89.60	84.80	63.60	86.80	91.20	89.20	76.80	10.80	85.96
R_λ	T=0	95.60	91.60	86.40	86.80	88.00	81.60	53.60	73.20	88.40	84.00	59.60	6.80	80.80
R_λ	T=0.6	96.40	92.00	86.40	87.60	89.20	80.40	50.40	74.40	90.00	88.00	58.00	8.00	81.16
R_λ	rep=1.05	96.80	92.80	85.60	87.60	87.60	81.20	58.80	76.00	90.80	87.60	59.20	10.00	82.18
$\mathrm{R}_{\mathrm{cov}}$	T=0	95.60	92.00	86.80	87.60	88.40	84.00	57.20	86.80	90.80	90.00	77.20	9.20	85.13
$\mathrm{R}_{\mathrm{cov}}$	T=0.6	96.00	92.40	84.80	87.60	88.80	84.80	61.20	86.00	92.00	89.20	77.60	9.60	85.49
$\mathrm{R}_{\mathrm{cov}}$	rep=1.05	94.40	93.20	88.80	88.00	88.40	86.00	60.80	84.80	92.00	91.20	78.40	10.80	86.00

Anchor is the full method. Its per-language results appear in the Anchor (Instruct initialization) rows of Table 6. Excluding bo, the Anchor–R4 Anchoring gain is $+1.31$ under greedy decoding and $+0.87$ under rep=1.05; bo remains a separate diagnostic. Thus, the gain varies with decoding. Repetition-penalty decoding also tends to improve the Tibetan result. For example, the R4 step-150 result rises from 6.80 to 11.60.

Paired significance tests. We test the six core ablation pairs and add R4–R1 and Anchor–R1, which correspond to the trade-off discussion in Section 3.2. The eight comparisons use the step-150 checkpoint and greedy decoding. They pair 2,750 questions across the original 11 languages, matched by language and question. We use paired bootstrap and exact McNemar tests. All confirmatory tests in the paper form three disjoint Benjamini–Hochberg (BH) families: the eight MGSM ablation comparisons in Table 8, the six anchoring-hyperparameter comparisons in Table 9, and the eight Lite-benchmark comparisons in Table 19. The resource-tier and per-language tests below are exploratory and lie outside these families.

The four comparisons that involve Anchoring remain significant after BH correction: Anchor against R1, R1b, and R4, and the decrease from Anchor to Rα. In every case, $p_{\mathrm{BH}} \leq 0.021$. The lower

confidence bound for Anchor–R4 is $+0.40$. The other four comparisons are not significant. We therefore detect no accuracy difference attributable to filtering or to the individual reward components under this evaluation. The confidence interval for R4 minus R1 is $[-0.73, +0.98]$. It includes zero and negative values. Thus, this comparison detects no significant accuracy cost, but it does not prove that the cost is exactly zero.

Table 8: Paired significance tests for the ablations on the original 11 MGSM languages (2,750 matched questions; baseline to candidate). Δ is the candidate-minus-baseline accuracy difference in percentage points. CIs use 10,000 paired-bootstrap samples. p is from the exact McNemar test, and p_{BH} is corrected within the family of eight comparisons.

Baseline → candidate	Δ	**95% CI**	p	$p_{\mathbf{BH}}$
R1 → R1b	+0.11	[-0.76, +0.98]	0.868	0.868
R1 → R4	+0.15	[-0.73, +0.98]	0.803	0.868
R2 → R4	+0.44	[-0.36, +1.27]	0.335	0.536
R3 → R4	+0.18	[-0.58, +0.98]	0.714	0.868
R1 → Anchor	**+1.45**	**[+0.58, +2.36]**	**0.002**	**0.015**
R1b → Anchor	**+1.35**	**[+0.44, +2.25]**	**0.005**	**0.016**
R4 → Anchor	**+1.31**	**[+0.40, +2.22]**	**0.006**	**0.016**
Anchor → Rα	**-1.05**	**[-1.85, -0.29]**	**0.011**	**0.021**

Resource-tier tests pair 1,750 questions from seven high-resource languages and 1,000 questions from four low-resource languages. These tests are exploratory and are not included in the main BH family. The Anchor-minus-R4 gain is significant in the low-resource group ($+2.10$, 95% CI $[+0.30, +3.90]$, $p = 0.028$) but not in the high-resource group ($+0.86$, $[-0.17, +1.89]$, $p = 0.119$); this does not establish a between-tier difference. No individual language remains significant after correcting the 11 McNemar tests within each comparison. Before correction, only en for R1b → Anchor ($p = 0.035$), and te for Anchor → Rα ($p = 0.021$) and R1 → Anchor ($p = 0.035$), are below 0.05. This indicates that the Anchoring gain is distributed across languages. Section 3.5 discusses the Anchoring dynamics.

Table 9: Paired tests for Anchoring hyperparameters on the original 11 languages (candidate minus Anchor, step 150). Δ is the language-macro-average accuracy difference in percentage points. CIs use language-stratified paired bootstrap. p_{BH} is corrected within six comparisons.

Configuration	**Decoding**	Δ	**95% CI**	$p_{\mathbf{BH}}$
R_λ	T=0	-5.27	[-6.51, -4.07]	¡0.001
R_λ	T=0.6	-4.47	[-5.71, -3.31]	¡0.001
R_λ	rep=1.05	-4.18	[-5.31, -3.05]	¡0.001
R_{cov}	T=0	-0.95	[-1.75, -0.15]	0.040
R_{cov}	T=0.6	-0.15	[-1.02, +0.73]	0.805
R_{cov}	rep=1.05	-0.36	[-1.20, +0.47]	0.532

Increasing the Anchoring weight tenfold makes all three decoding results significantly lower than Anchor. Their confidence intervals exclude zero and $p_{\mathrm{BH}} < 0.001$. Extending Anchoring coverage to the whole reasoning segment reduces greedy accuracy by 0.95 points ($p_{\mathrm{BH}} = 0.040$); its sampled-decoding contrasts are inconclusive, and for those two settings the correct conclusion is that no stable difference is detected, not that the method has been proved insensitive to coverage.

C.3 Full Language-Behavior Results (A and M)

This section gives the per-language composition and correct-only aggregation of the A and M columns in Tables 1, 2, and 3. The signal definitions match training (Section 2.3). We first average within each language and then macro-average across languages. A averages the original 11 MGSM languages; M averages their ten non-English languages.

Within each language, all evaluated responses form the denominator: no valid public chunk gives $A = 0$, and fewer than two valid reasoning chunks gives $M = 0$. The corresponding ineligible rates appear in Table 14. Tibetan (bo) is a separate per-language diagnostic and is reported only in the

per-language tables. Correct-only columns similarly use all correct responses within each language, including those without valid chunks. Reasoning-language shares p_g, p_{en}, p_o remain conditional on at least one valid reasoning chunk. Tables 12 and 13 give correct-only results for the gated interpretation of how the model expresses itself once correct.

For compactness in Tables 10–14, "Qwen3-8B Inst." denotes the Qwen3-8B-Instruct initialization. "Anchor-SFT$_{5e-6}$" and "Anchor-SFT$_{1e-6}$" denote Anchor initialized from $R_{\mathrm{SFT}}(5 \times 10^{-6})$ and $R_{\mathrm{SFT}}(1 \times 10^{-6})$, respectively. Anchor$_{80}$ uses the Instruct initialization at step 80, while "Anchor" uses the same initialization at step 150.

Table 10: Public alignment A (%) with all-response denominators; missing valid public chunks score zero. Macro avg. excludes bo.

Model	en	es	fr	de	zh	ja	sw	bn	ru	th	te	bo	Macro avg.
Qwen3-8B Inst.	57.2	75.3	75.3	61.3	49.4	64.9	8.0	23.4	43.0	26.0	55.4	3.7	49.0
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	80.8	67.6	64.2	64.5	29.9	56.0	16.4	46.7	62.6	64.8	48.0	16.8	54.7
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	66.1	60.3	73.4	56.9	27.3	47.2	5.2	18.4	45.9	16.0	23.1	6.0	40.0
Anchor-SFT$_{5e-6}$	74.0	58.9	60.6	57.2	21.5	52.4	20.3	44.8	57.2	54.0	48.0	34.0	49.9
Anchor-SFT$_{1e-6}$	62.3	69.4	73.1	63.4	33.0	54.8	11.6	37.2	57.5	29.6	40.0	13.2	48.4
Anchor$_{80}$	71.1	80.5	84.8	80.4	66.2	85.3	21.9	60.6	81.4	60.8	79.6	12.8	70.2
Anchor	73.4	80.6	86.7	83.3	72.1	89.6	26.2	67.8	85.0	68.7	82.6	15.2	74.2
R1	59.2	77.3	77.7	67.6	53.5	71.5	12.0	29.7	53.6	34.0	63.2	6.1	54.5
R1b	69.1	79.8	83.8	78.5	64.2	81.8	18.7	57.2	75.1	60.4	77.1	10.5	67.8
R2	68.0	77.3	84.2	78.3	64.0	82.7	21.0	55.2	75.3	58.7	77.7	11.2	67.5
R3	70.4	79.4	84.5	78.6	68.2	82.8	20.4	59.1	79.1	65.9	77.0	11.8	69.6
R4	71.5	81.8	84.6	80.9	64.6	83.0	20.8	57.7	76.8	61.4	76.6	10.8	69.1
Rα	72.7	80.9	85.4	80.7	68.6	88.9	22.4	69.5	84.3	68.7	87.0	19.1	73.5
R$_\lambda$	76.7	78.6	86.0	66.6	66.0	85.8	19.0	44.0	62.9	65.2	62.8	16.4	64.9
R$_{\mathrm{cov}}$	73.3	81.9	85.8	81.5	67.3	89.2	22.8	66.7	81.3	66.7	83.6	15.6	72.7

Table 11: Reasoning mixing M (%) with all-response denominators; missing valid reasoning chunks score zero. English and bo are excluded from Macro avg.

Model	es	fr	de	zh	ja	sw	bn	ru	th	te	bo	Macro avg.
Qwen3-8B Inst.	8.2	9.2	10.3	2.1	5.7	5.2	9.7	8.1	8.9	12.6	8.7	8.0
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	6.5	5.9	7.4	2.3	4.0	2.9	3.9	4.7	6.3	6.3	4.1	5.0
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	5.2	7.6	8.1	2.2	5.0	3.4	4.5	5.4	6.8	5.6	3.3	5.4
Anchor-SFT$_{5e-6}$	4.2	4.9	5.9	1.1	2.7	2.3	3.0	4.8	4.6	3.7	4.2	3.7
Anchor-SFT$_{1e-6}$	6.8	6.3	8.6	1.7	4.2	2.5	3.1	4.5	4.3	5.1	1.9	4.7
Anchor$_{80}$	9.1	10.0	10.2	2.4	6.7	6.1	11.3	10.2	12.8	15.1	13.3	9.4
Anchor	9.5	12.6	10.8	2.5	6.5	7.3	11.8	9.5	13.9	15.9	21.6	10.0
R1	9.2	10.5	9.5	2.2	5.7	5.5	9.3	7.5	9.3	11.5	9.4	8.0
R1b	8.5	9.4	8.7	2.1	5.5	5.2	10.3	7.5	10.0	13.2	11.2	8.0
R2	8.1	10.0	8.8	1.9	5.6	5.8	10.5	8.0	10.7	13.2	11.9	8.3
R3	8.2	10.0	8.6	2.3	5.9	4.8	10.1	7.5	9.1	12.9	11.9	7.9
R4	8.7	9.3	8.9	2.3	5.9	5.5	9.2	7.7	10.0	12.7	10.1	8.0
Rα	9.6	12.2	11.1	2.2	7.4	6.9	12.7	9.1	13.1	16.8	22.2	10.1
R$_\lambda$	4.0	17.0	5.6	1.8	0.0	10.0	0.0	0.0	0.4	0.4	23.5	3.9
R$_{\mathrm{cov}}$	8.8	13.4	11.3	2.5	7.7	7.1	14.4	10.3	13.7	16.3	19.2	10.6

C.4 Composition of p_o

To further analyze the multilingual composition of the reasoning segment, we define $p_o = 1 - p_g - p_{en}$ as the residual fastText probability mass after excluding the current question language and English on a per-response basis. Therefore, p_o does not correspond to any fixed "third language": the same language contributes to p_g when it is itself the question language, but may contribute to p_o

Table 12: Public alignment A (%) on correct responses, using every correct response as the denominator. Macro avg. excludes bo.

Model	en	es	fr	de	zh	ja	sw	bn	ru	th	te	bo	Macro avg.
Qwen3-8B Inst.	59.0	78.7	81.0	64.6	53.5	70.7	10.9	26.0	45.8	28.7	63.1	10.5	52.9
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	89.0	76.0	75.9	69.0	32.5	60.0	33.6	61.2	71.0	81.1	72.1	66.5	65.6
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	83.2	77.9	88.1	74.7	43.9	72.1	10.0	27.3	54.1	15.5	36.0	0.0	53.0
Anchor-SFT$_{5e-6}$	87.4	77.0	74.3	72.7	28.9	66.1	39.6	64.4	65.2	73.3	76.0	99.8	65.9
Anchor-SFT$_{1e-6}$	71.3	81.3	87.8	81.2	44.7	71.1	24.6	59.0	66.2	41.1	60.6	66.6	62.6
Anchor$_{80}$	73.8	82.9	88.8	82.6	69.0	92.6	27.8	64.9	85.2	63.7	91.2	37.5	74.8
Anchor	75.1	82.3	90.1	85.0	72.9	94.0	30.7	73.0	87.7	72.1	92.8	23.8	77.8
R1	61.9	80.9	82.6	72.3	59.1	78.5	14.8	32.1	56.3	35.7	74.6	17.6	59.0
R1b	71.8	83.2	88.6	82.2	69.0	90.2	24.7	60.5	78.1	62.6	89.2	21.0	72.7
R2	70.9	82.0	88.1	83.4	70.0	89.9	26.4	61.8	80.8	62.4	89.8	22.2	73.2
R3	73.0	82.9	88.3	80.7	71.0	91.0	26.6	63.8	82.8	70.3	87.9	33.3	74.4
R4	73.4	83.8	89.0	83.7	68.3	91.2	26.4	63.9	79.6	66.0	90.1	23.5	74.1
Rα	74.5	82.7	89.0	83.1	71.1	93.9	26.0	72.8	86.9	70.6	97.1	38.9	77.1
R$_\lambda$	78.3	80.8	88.5	67.8	66.5	90.4	24.1	50.2	65.7	70.9	91.3	23.5	70.4
R$_{\mathrm{cov}}$	74.8	84.2	89.2	83.1	69.9	93.4	27.3	71.8	84.4	69.7	94.8	30.4	76.6

Table 13: Reasoning mixing M (%) on correct responses, using every correct response as the denominator. English and bo are excluded from Macro avg.

Model	es	fr	de	zh	ja	sw	bn	ru	th	te	bo	Macro avg.
Qwen3-8B Inst.	7.5	8.6	10.5	2.1	6.0	6.2	9.9	8.1	9.7	14.0	12.7	8.3
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	6.8	7.4	8.3	2.5	4.2	6.0	4.6	4.8	5.7	9.6	17.6	6.0
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	5.4	8.6	8.7	2.7	4.6	9.5	6.0	4.3	9.0	7.0	12.4	6.6
Anchor-SFT$_{5e-6}$	5.1	4.6	6.5	1.5	3.6	6.0	6.6	5.6	5.7	5.6	29.0	5.1
Anchor-SFT$_{1e-6}$	8.0	7.8	11.5	2.0	5.5	4.9	4.1	5.4	6.0	10.5	16.7	6.6
Anchor$_{80}$	8.2	9.7	9.5	2.2	6.5	6.6	11.7	9.7	12.9	14.4	28.4	9.1
Anchor	8.7	12.2	10.3	2.3	6.3	7.7	11.8	9.1	13.7	16.3	43.1	9.8
R1	9.0	10.3	9.4	2.3	5.9	6.3	9.9	7.6	9.3	12.9	25.6	8.3
R1b	8.2	9.2	8.0	2.2	5.5	5.6	10.0	7.4	10.2	13.5	28.8	8.0
R2	8.2	9.8	8.9	2.2	5.6	6.6	11.0	7.9	11.1	14.4	28.9	8.6
R3	7.9	9.2	8.0	2.2	6.3	6.5	10.3	7.4	9.8	13.0	24.1	8.1
R4	8.0	9.2	8.3	2.3	6.4	6.0	10.0	7.6	10.7	14.2	17.9	8.3
Rα	8.5	12.1	10.8	2.1	7.6	7.3	11.1	8.0	13.3	15.1	39.3	9.6
R$_\lambda$	4.2	17.4	5.7	1.8	0.0	11.5	0.0	0.0	0.5	0.6	48.1	4.2
R$_{\mathrm{cov}}$	7.9	13.2	10.4	2.5	7.2	8.6	14.1	10.1	14.2	16.5	36.5	10.5

Table 14: MGSM language-behavior summary (%) over the original 11 languages (A) or ten non-English languages (M and reasoning-language shares). A/M use all responses, with zero for responses lacking valid chunks; p_g, p_{en}, p_o condition on valid reasoning chunks. bo is reported only in the per-language tables.

Model	A	M	A correct-only	M correct-only	A unparseable	M unparseable	Empty public	p_g	p_{en}	p_o
Qwen3-8B Inst.	49.0	8.0	52.9	8.3	45.0	9.0	43.5	5.7	87.7	6.6
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	54.7	5.0	65.6	6.0	39.3	23.3	22.3	2.7	89.5	7.8
$R_{\mathrm{SFT}}(1 \times 10^{-6})$	40.0	5.4	53.0	6.6	54.5	25.4	48.3	5.3	86.9	7.8
Anchor-SFT$_{5e-6}$	49.9	3.7	65.9	5.1	44.6	29.6	28.5	10.9	82.3	6.8
Anchor-SFT$_{1e-6}$	48.4	4.7	62.6	6.6	47.9	28.4	40.6	4.1	88.1	7.8
Anchor$_{80}$	70.2	9.4	74.8	9.1	22.0	6.5	19.6	5.3	88.0	6.7
Anchor	74.2	10.0	77.8	9.8	17.6	5.4	15.2	7.5	85.9	6.6
R1	54.5	8.0	59.0	8.3	39.1	9.0	36.9	5.7	87.7	6.6
R1b	67.8	8.0	72.7	8.0	24.7	7.8	22.5	8.4	85.2	6.4
R2	67.5	8.3	73.2	8.6	24.8	8.9	22.5	7.9	85.6	6.4
R3	69.6	7.9	74.4	8.1	22.4	8.4	20.4	7.5	86.1	6.5
R4	69.1	8.0	74.1	8.3	23.3	8.2	20.9	7.1	86.5	6.5
Rα	73.5	10.1	77.1	9.6	18.5	5.5	16.7	7.6	85.8	6.7
R$_\lambda$	64.9	3.9	70.4	4.2	28.0	14.7	24.5	87.1	9.5	3.4
R$_{\mathrm{cov}}$	72.7	10.6	76.6	10.5	19.4	5.9	17.1	8.5	85.0	6.6

Table 15: Per-language accuracy (%) on Global-MMLU-Lite under T=0.

Model	ar	bn	cs	cy	de	en	es	fr	hi	hu	id	it
Qwen3-8B-Instruct	76.25	70.50	75.00	58.00	79.25	82.00	78.50	75.75	69.50	79.25	78.25	78.25
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	50.00	38.75	53.25	33.50	62.75	63.00	59.75	60.00	41.50	48.25	56.25	59.50
R1	74.75	70.50	78.00	59.75	80.00	82.75	80.25	79.00	71.00	78.00	78.75	80.00
R4	74.25	69.00	77.75	61.50	77.50	81.00	80.00	78.00	72.50	75.00	81.25	79.75
Anchor	70.75	68.75	76.75	60.75	79.50	83.00	78.50	78.25	73.00	76.75	78.25	78.75

Model	ja	ko	my	or	pt	sk	sq	sw	tg	yo	zh	Macro avg.
Qwen3-8B-Instruct	78.75	77.25	57.14	58.75	78.50	76.50	73.75	47.50	57.25	30.50	79.00	70.23
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	56.00	51.50	23.06	16.00	62.25	53.25	48.25	24.75	26.50	15.50	61.75	46.32
R1	78.50	76.00	57.64	58.25	80.50	75.00	76.50	48.00	57.25	31.25	79.00	70.90
R4	77.25	75.00	60.40	58.25	78.75	78.00	72.25	48.25	62.75	30.00	79.00	70.76
Anchor	79.75	75.25	54.14	57.25	78.00	77.75	72.75	47.50	60.00	31.50	80.25	70.31

for questions written in other languages. In Anchor's Lite evaluations, p_o is approximately 7.3% on Global-MMLU-Lite and 8.8% on MMLU-ProX-Lite. Further decomposition of the full fastText label-probability distribution shows that this mass is spread across multiple languages. These include primarily Western European languages such as Spanish, French, German, Portuguese, and Italian; Eastern European languages such as Russian and Romanian; and South and Southeast Asian languages such as Telugu, Bengali, Vietnamese, Hindi, Thai, and Indonesian. The share attributable to any individual language is small. Even the most frequent labels account for only part of p_o, while the remaining probability mass is dispersed across a large number of low-probability language labels, forming a clear long-tail distribution. Therefore, p_o is better interpreted as broad multilingual participation retained during reasoning rather than as concentrated drift toward a particular additional language. Note that these statistics are computed over responses containing at least one valid reasoning chunk and measure fastText label probability mass rather than token counts or hard language-classification proportions.

D Full Results on the Two Lite Benchmarks

This section gives the per-language results, language-behavior metrics, and paired significance tests for Figure 5 in Section 3.6. Appendix A.5 gives the dataset sizes and decoding settings. Accuracy in the per-language tables is the macro-average under T=0, in percent. Grading uses the final `####` `<option letter>` marker. The language-behavior metrics follow Appendix C.3. Their ProX-Lite macro-averages exclude wo and zu because fastText does not support them.

Table 16: Per-language accuracy (%) on MMLU-ProX-Lite under T=0.

Model	af	ar	bn	cs	de	en	es	fr	hi	hu
Qwen3-8B-Instruct	56.97	53.91	53.74	55.27	56.63	55.61	58.50	58.33	51.02	57.31
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	25.34	23.81	17.86	28.91	28.91	31.29	33.67	33.33	16.16	21.94
R1	60.03	54.59	53.40	57.48	57.31	58.16	59.52	59.69	53.23	58.84
R4	57.48	54.42	53.57	59.86	58.33	57.82	60.03	60.71	56.46	59.35
Anchor	60.88	58.67	55.10	60.71	61.05	57.82	60.88	60.88	57.14	60.54

Model	id	it	ja	ko	mr	ne	pt	ru	sr	sw
Qwen3-8B-Instruct	58.16	58.84	53.23	53.40	56.29	52.89	58.33	56.46	57.99	36.73
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	29.59	33.84	31.12	28.40	17.18	16.16	34.86	33.67	29.25	11.90
R1	59.18	59.35	54.93	54.42	57.14	54.42	59.35	59.69	57.65	38.10
R4	59.35	60.71	57.65	55.95	54.93	55.27	59.69	60.20	59.18	36.05
Anchor	61.90	63.44	56.97	56.12	56.63	57.65	61.22	58.84	60.20	40.99

Model	te	th	uk	ur	vi	wo	yo	zh	zu	Macro avg.
Qwen3-8B-Instruct	50.17	54.76	57.14	55.10	57.99	26.70	25.17	55.61	26.02	52.01
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	9.35	24.83	28.23	16.67	30.44	7.82	7.99	32.99	5.78	23.84
R1	54.25	56.80	57.48	57.14	59.52	29.08	23.81	56.46	29.59	53.47
R4	54.25	55.78	57.31	54.93	61.22	28.91	26.19	55.61	27.38	53.75
Anchor	56.29	56.63	57.48	59.01	59.35	31.80	26.87	58.67	29.42	55.28

Table 17: Language-behavior metrics (%) under rep=1.05. A/M use all-response denominators, with zero for missing valid chunks; p_g, p_{en}, p_o condition on valid reasoning chunks. ProX-Lite excludes unsupported wo and zu.

Model	Benchmark	A	M	p_g	p_{en}	p_o
Qwen3-8B-Instruct	Global-MMLU-Lite	1.2	5.5	2.4	90.7	6.9
Qwen3-8B-Instruct	MMLU-ProX-Lite	1.4	4.5	2.4	89.1	8.5
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	Global-MMLU-Lite	72.0	8.0	3.5	88.9	7.6
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	MMLU-ProX-Lite	44.6	6.1	2.8	88.3	8.8
R1	Global-MMLU-Lite	0.2	5.8	2.3	90.7	7.0
R1	MMLU-ProX-Lite	0.3	4.5	2.4	89.0	8.6
R4	Global-MMLU-Lite	0.9	5.6	2.4	90.6	7.0
R4	MMLU-ProX-Lite	1.0	4.7	2.6	88.8	8.6
Anchor	Global-MMLU-Lite	2.2	8.4	3.7	89.1	7.3
Anchor	MMLU-ProX-Lite	1.9	6.4	5.2	86.0	8.8

On multiple-choice tasks most models emit a bare answer marker, so A is small for all models except R_SFT(Table 18),

The directional pattern agrees with MGSM. M changes little from R1 to R4 (5.8/4.5 on Global-MMLU-Lite/MMLU-ProX-Lite for R1 and 5.6/4.7 for R4); for Anchor it rises to 8.4 and 6.4. The reasoning-language shares remain conditional on valid reasoning chunks. These are descriptive fastText proxies, and we make no significance claim about them.

For Anchor–R1 on MMLU-ProX-Lite, 22 of 29 languages improve. The largest gains are $+4.08$ for ar and it, $+3.91$ for hi, and $+3.74$ for de; the largest decrease is -0.85 for ru. On Global-MMLU-Lite, eight languages improve and 15 decline. These tests are paired diagnostics from one run and do not estimate variance across seeds.

Table 18: Computable proportions (%) for A, M, and the vocabulary-share distribution (p_g, p_{en}, p_o), together with the empty-public-segment rate.

Model	Benchmark	A comp.	M comp.	$p_\cdot$ comp.	Empty public
Qwen3-8B-Instruct	Global-MMLU-Lite	1.5	97.6	97.6	98.5
Qwen3-8B-Instruct	MMLU-ProX-Lite	2.3	78.1	78.1	97.6
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	Global-MMLU-Lite	79.2	87.0	87.0	14.2
$R_{\mathrm{SFT}}(5 \times 10^{-6})$	MMLU-ProX-Lite	52.6	65.0	65.0	38.8
R1	Global-MMLU-Lite	0.3	97.8	97.8	99.7
R1	MMLU-ProX-Lite	0.5	78.8	78.8	99.5
R4	Global-MMLU-Lite	1.0	98.0	98.0	98.9
R4	MMLU-ProX-Lite	1.4	79.8	79.8	98.6
Anchor	Global-MMLU-Lite	2.5	98.6	98.7	97.5
Anchor	MMLU-ProX-Lite	2.5	82.9	82.9	97.5

Table 19: Paired significance tests under T=0. Δ is candidate-minus-baseline macro-average accuracy in percentage points. CIs use language-stratified paired bootstrap. p_{BH} is corrected within eight comparisons. Bold marks significance after correction.

Comparison	Global-MMLU-Lite Δ [95% CI],p_{BH}	MMLU-ProX-Lite Δ [95% CI],p_{BH}
Anchor - Instruct	+0.08 [-0.63, +0.77], 0.858	**+3.27 [+2.73, +3.80], ¡0.001**
Anchor - R1	-0.59 [-1.29, +0.11], 0.222	**+1.81 [+1.29, +2.33], ¡0.001**
Anchor - R4	-0.45 [-1.14, +0.26], 0.357	**+1.54 [+1.01, +2.05], ¡0.001**
R4 - R1	-0.14 [-0.83, +0.57], 0.816	+0.28 [-0.23, +0.77], 0.397

E Qualitative Case Analysis

Setup. We select five MGSM questions for each of three low-resource languages—Tibetan (bo, from our Tibetan extension), Bengali (bn), and Telugu (te)—and audit the responses of three closed-source models (GPT-5.5, Gemini-3.5-Flash, and DeepSeek-V4-Flash) and Anchor under the MGSM evaluation prompt template. Figures 6–8 show all 15 cases; reasoning segments are truncated for display, and public segments are shown in full. GPT-5.5 and Gemini-3.5-Flash are constrained by their respective providers, and neither was able to provide the reasoning process.

GPT-5.5: answer-only degeneration. GPT-5.5 returns the correct value on 14 of 15 questions, but every response collapses to an answer-only output: the entire response is at most one short lead-in phrase followed by the answer marker. The question-language audience receives no explanation at all. One response additionally emits two answer markers, the second written in Tibetan numerals.

Gemini-3.5-Flash: over-localized answers. Gemini-3.5-Flash writes open step-by-step solutions, and only 3 of 15 cases are fully correct. Its dominant failure is over-localization of the answer itself: in 7 of 15 cases the final answer is written in the numeral system of the question language—Tibetan digits (e.g., the Tibetan numeral for 70), Tibetan number words (sixty), or Bengali digits (e.g., the Bengali numeral for 51)—so the numeric answer cannot be parsed. The remaining cases include two calculation errors, one of them combined with a duplicated answer marker; one response whose public segment is written entirely in English; and two answer-only outputs.

DeepSeek-V4-Flash: English-locked reasoning. DeepSeek-V4-Flash is the only closed-source model that reliably produces both segments (14 of 15), and 4 of 15 cases are fully correct. Its reasoning, however, is locked to English: in 8 of 15 cases the reasoning segment contains no question-language text at all, frequently opening with an explicit translation move such as "Let me work through the problem in English," and in 4 of these cases the public segment also degenerates to a bare answer marker. The remaining errors are two calculation errors (one response reports the unrounded value 106.12 where the gold answer is 106) and one response without a reasoning segment.

Anchor. Anchor is fully correct on all 15 cases: every response gives the correct value and a step-by-step public explanation in the question language, and every reasoning segment mixes the question language with English (2–19% of reasoning-script characters, against zero in the 8 English-locked DeepSeek-V4-Flash cases).

Summary. In this audit, each closed-source model fails a different requirement: GPT-5.5 omits the public explanation entirely, Gemini-3.5-Flash localizes the surface form of the answer at the cost of parseability and accuracy, and DeepSeek-V4-Flash serves the public segment in the question language while reasoning only in English. None of the three produces, on all 15 cases, the combination that Anchor produces: correct values, question-language public explanations, and reasoning that involves the question language.

Case study: Tibetan (bo) — 5 questions

Language	Question	GPT-5.5	Analysis	Gemini-3.5-Flash	Analysis	DeepSeek-V4-Flash	Analysis	Anchor	Analysis
Tibetan (bo)	Jason གཙུབ་ཀྱིས་ 1,000 ཐོག་བཀོ་ཤོག་གངས་ཅན་ཞིག་ཐོབ་པ་ཡིན། ཁོས་ཀྱིས་ཁ་པར་ཁང་གི་ཉིན་རེའི་ཁ་པར་ 15 བཀོ་བ་ཡོད་པ་དང་། ཁ་པར་ཁང་གིས་ཚེས་རེ་ལ་ 30 ཐོག་ཁྲལ་ལེན་ཡོད། Jason ཀྱིས་ཟླ་རེ་ལ་ 15-minute call with his boss, and he's had 300 extra minutes of call this month to other people. How many minutes does Jason have left if this month has 30 days?	Jason ཀྱིས་ཟླ་བའི་ནང་དུ་སྐར་མ་ ### 250	Missing public explanation	1. **Total monthly allowance**: 1,000 minutes. 2. **Minutes used for daily calls with his sponsor**: • 15 minutes/day × 30 days = 450 minutes. 3. **Extra minutes used with other people**: 300 minutes. 4. **Total minutes used**: • 450 minutes + 300 minutes = 750 minutes. 5. **Remaining minutes**: • 1,000 minutes - 750 minutes = 250 minutes. Jason ཀྱིས་ཟླ་བའི་ནང་དུ་ 250 སྐར་མ། #### 250	Public segment in English	<think>Let me work through the problem in English: Jason has a monthly phone plan with 1,000 minutes. He must call his sponsor 15 minutes every day. This month, he's spent 300 extra minutes talking to other people. This month has 30 days. First, calculate the minutes used for calling his sponsor: 15 ·(truncated)· </think> ### 250	Missing public explanation and reasoning not in question language	<think>Okay, let's try to solve this problem step by step. So, first, I need to understand what's being asked here. The problem is in Tibetan, so I'll need to parse the information carefully. Billy, so, the problem says: Jason has a debt of 1,000 yen. Each day, he can pay 15 yen. This month, he has paid 300 yen more than usual. The month has 30 days. How much is Jason ·(truncated)· </think> #### 250	Fully correct
Tibetan (bo)	Killian ཀཻལིཡཱན་ཞེས་པའི་ཁྱེད་ཀྱི་ཟླ་བའི་ནང་དུ་ཁྱེད་ཀྱིས་ཁ་པར་ 20 བཀོ་ཞིང་། ཁྱེད་ཀྱིས་ཁ་པར་ཁང་གི་ཟས་ཐོ་ཡིན་ཏེ་ཁྱེད་ཀྱིས་ཟས་ཐོ་ཡི་ནང་ནས་ཟས་རིགས་ 2 ཐོབ་པ་ཡིན། EN: Killian has two recipes for preparing dishes, one having 20 instructions and the second one having twice as many instructions as the first one. How many instructions does Killian have to read to prepare the two dishes?	### 60	Missing public explanation	#### ༦༠	No public segment and no numerical output	<think>The problem is in Tibetan. Let me translate it mentally: "Killian has two menus. The first one contains 20 items. The number of items in the second menu is twice the first. To make both menus, how many items does Killian need to read?" So first menu = 20 items. Second menu = 2 × 20 = 40 ·(truncated)· </think> ### 60	Missing public explanation and reasoning not in question language	<think>Okay, let's try to solve this problem step by step. So, the problem is about Killian having two boxes. The first box has 20 items, and the second box's items are twice the number of the first box. Wait, no, wait. Wait, the problem says the second box's quantity is equal to the first box's double. ·(truncated)· </think> #### 60	Fully correct
Tibetan (bo)	Carmen ཅཱར་མེན་གྱིས་ Sudoku རྩེ་དཔྱད་ 8 ཞིབ་བསྡུར་བྱས་ཏེ་ 10 དཀྱིལ་འཁོར་ཞིག་ཐོབ་པ་ཡིན། ཁོས་ཀྱིས་ Sudoku 3 དང་ Sudoku རྩེ་དཔྱད་ 8 ཞིབ་བསྡུར་བྱས་ཏེ་ 3 དང་ Sudoku རྩེ་དཔྱད་ 8 ཞིབ་བསྡུར་བྱས་ཏེ་ 5 དཀྱིལ་འཁོར་ཞིག་ཐོབ་པ་ཡིན། ཁོས་ཀྱིས་ Sudoku རྩེ་དཔྱད་ 8 ཞིབ་བསྡུར་བྱས་ཏེ་ 3 དང་ Sudoku རྩེ་དཔྱད་ 8 ཞིབ་བསྡུར་བྱས་ཏེ་ 5 དཀྱིལ་འཁོར་ཞིག་ཐོབ་པ་ཡིན། EN: It takes Carmen 10 minutes to finish a sudoku puzzle and 5 minutes to finish a sudoku puzzle. Over the weekend she solved 3 crossword puzzles and 8 sudoku puzzles. How much time did she spend playing these games?	### 70	Missing public explanation	1. Carmen ཅཱར་མེན་གྱིས་ 3 ཞིབ་བསྡུར་བྱས་ཏེ་ $3 times 10 = 30$ དཀྱིལ་འཁོར། 2. ཁོ་ Sudoku རྩེ་དཔྱད་ 8 ཞིབ་བསྡུར་བྱས་ $8) times 5 = 40$ དཀྱིལ་འཁོར། 3. རྩིས་བཀྲལ་ $30 + 40 = 70$ སྐར་མ། #### ༧༠	Final answer not numeric	<think> Carmen spends 10 minutes per word puzzle and 5 minutes per Sudoku. She completed 3 word puzzles and 8 Sudokus. Total time = 3×10 + 8×5 = 30 + 40 = 70 minutes.</think> #### 70	Reasoning not in question language	<think>Okay, let's try to solve this problem step by step. First, I need to understand what the question is asking. The problem is about Carmen and the time she spends on different activities. Let me break down the information given. The first part says: "Carmen ·(truncated)· </think> #### 70	Fully correct
Tibetan (bo)	Mandy ཡིས་ Benedict ལས་ 100 ཀྲམ་དོར་ཏེ་དངུལ་ 3 ཟླ་བའི་ནང་དུ་ 2% ཡིན་པའི་ཐོག་ནས་ Benedict ལ་ $100 ཕྱིར་ལོག་བྱེད་དགོས་པ་ཡིན། EN: Mandy owes Benedict $100. They agreed to have monthly interest of 2%. If Mandy was able to pay it after 3 months, how much should she give to Benedict?	### 106	Missing public explanation	#### ༡༠༦	Final answer not numeric	<think>Let me parse the Tibetan problem: "Mandy ཡིས་ Benedict ལས་ 100 ཀྲམ་དོར་ཏེ་ ... Mandy owes Benedict 100 dollars. The monthly interest rate is 2%. The question is: If they agreed that the monthly interest rate is 2%, and Mandy ... 3 months, how much should Mandy give to Benedict ·(truncated)· </think> Mandy ཡིས་ Benedict ལ་ 100 ཀྲམ་ རྩིས་ཏེ་ 2% ཡིན་པས་ ཟླ་ 3 རྗེས་སུ་ དངུལ་ཕོགས་ཚུལ་ $100 \times (1 + 0.02)^3 = 100 \times 1.061208$ = 106.1208 ཨེ། #### 106.12	Calculation error	<think>Okay, let me try to figure out this problem. So, Mandy has $100 to invest in Benedict. Both of them have an interest rate of 2% per month. Then, Mandy wants to withdraw some money after 3 months. The question is, how much does Benedict need to ·(truncated)· </think> Mandy ཡིས་ $100 ལས་ 2% ཡིན་པས་ ཟླ་ 3 རྗེས་སུ་ དངུལ་ཕོགས་ཚུལ་ $100 \times (1 + 0.02)^3 = 100 \times 1.061208$ = 106.1208 ཨེ། #### 106	Missing public explanation and reasoning not in question language	<think>Okay, let's try to figure out this problem step by step. So, the problem is about Jean's shopping and some calculations. Let me break down the information given. First, it says that Jean has 30 shopping items. Then, Jean used 2 of them. Wait, the original problem might be in Tibetan, but I need to ·(truncated)· </think> #### 14	Fully correct
Tibetan (bo)	Jean ཟངས་ཀྱིས་ 30 ཁྱེད་ Jean ཁྱེད་ཀྱིས་ 2 ཁྱེད་ 1 ཟངས་ལག་ཁྱེར

Case study: Bengali (bn) — questions 1–2

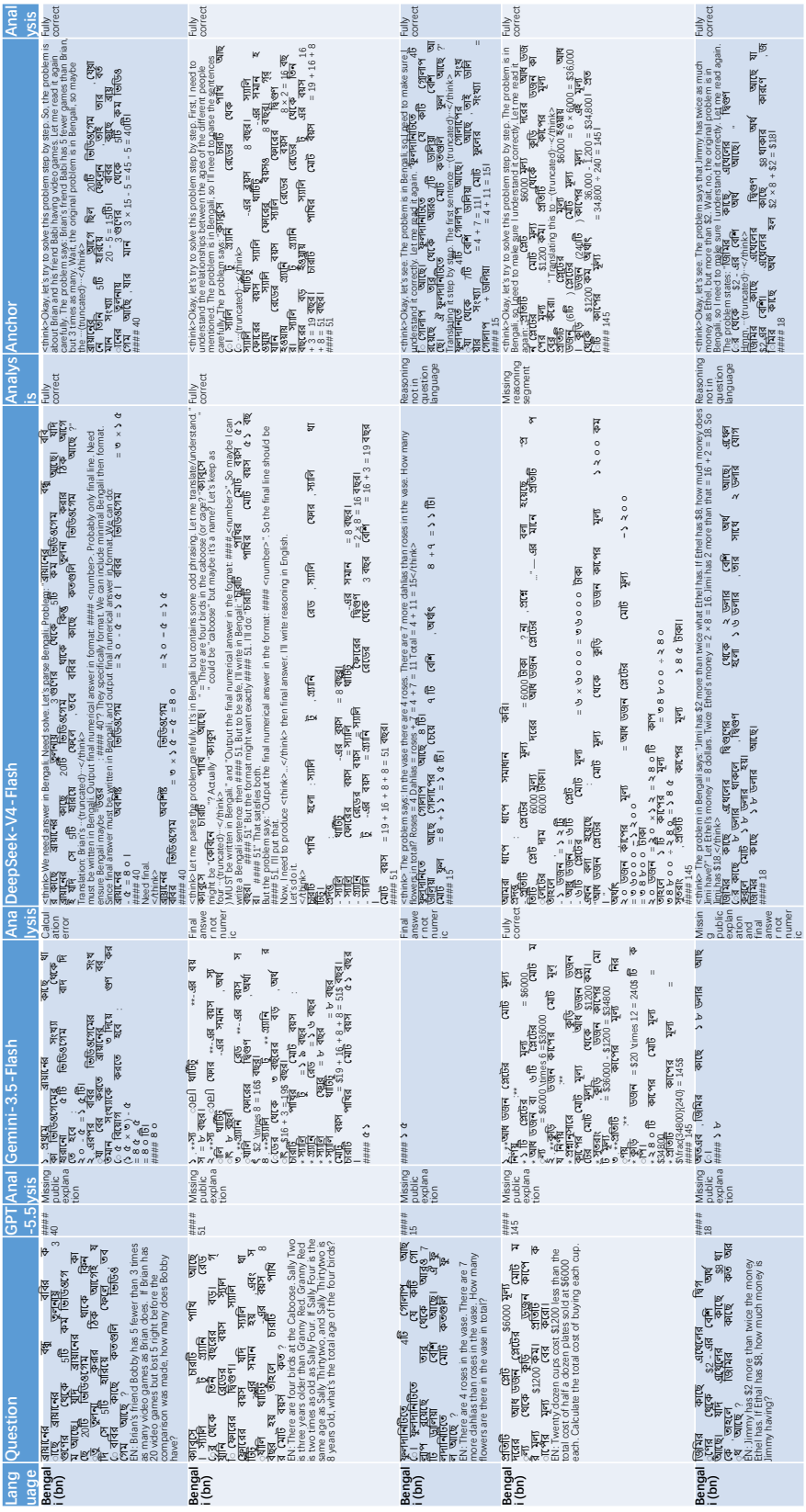

Figure 7: Qualitative case analysis for Bengali (bn).

Case study: Telugu (te) — questions 1–2



Figure 8: Qualitative case analysis for Telugu (te).