

Chain-of-Thought in a Low-Resource Language: A Factorial Study of Exemplar and Reasoning Language Effects in Urdu

Ateeb Ahmed

Saarland University

atah00001@stud.uni-saarland.de

Abstract

Chain-of-thought (CoT) prompting has been shown theoretically to extend the computational capacity of transformers by providing intermediate reasoning steps (Feng et al., 2023; Amiri et al., 2025). Whether these benefits carry over to low-resource languages is not well understood, and Urdu in particular has received very little attention in this context. We investigate two prompt factors (the language of few-shot exemplars and the language of the reasoning chain) using a 2×2 factorial design on MGSM Urdu and Math-500 Urdu, evaluated with Gemini 2.5 Flash. We find a significant interaction between exemplar and CoT language (item-level logistic regression, $z = 6.9$, $p < 0.001$): switching to Urdu CoT with English exemplars causes a 34.8 percentage point accuracy drop on MGSM, while the same switch with Urdu exemplars results in only a 1.6 point drop. We also find that Urdu CoT produces longer token sequences than English CoT (139 vs. 101 tokens on average) but not correspondingly more discrete reasoning steps: a script-aware, manually validated step count shows Urdu CoT yields slightly fewer discrete steps (2.5 vs. 3.0 on average), with the reduction concentrated in the mismatched-exemplar condition, indicating that added verbosity does not translate into added reasoning structure. Chain length correlates positively with accuracy ($r = 0.42$, $p < 0.001$), consistent with the theoretical framework. The main takeaway is that exemplar language is a critical moderator of CoT effectiveness in Urdu.

1 Introduction

Large language models (LLMs) equipped with chain-of-thought (CoT) prompting have shown strong improvements on reasoning tasks (Wei et al., 2022; Kojima et al., 2022), but the bulk of this research has been conducted in English. As LLMs get deployed across languages, it becomes important to understand whether CoT ben-

efits transfer, and what prompt-level factors determine how well they do.

The theoretical motivation is clear. Feng et al. (2023) and Amiri et al. (2025) argue that CoT works by giving the model extra computation through intermediate steps. Without CoT, a transformer is limited to problems solvable in a single pass. Amiri et al. (2025) prove lower bounds on how many steps are needed for certain problem classes. The implication is direct: if reasoning in a particular language causes the model to produce fewer or shorter reasoning steps, that is not just a stylistic difference; it is a reduction in computational capacity, with predictable consequences for accuracy.

This raises a natural question: when a model reasons in Urdu, does the structure of its reasoning chain change? Prior work on multilingual CoT has mostly shown that English CoT outperforms native-language CoT (Shi et al., 2022), with the gap depending on resource level (Huang et al., 2023; Barua et al., 2025). Ferrao et al. (2025) found that step counts vary across languages and correlate with accuracy, but only covered Bengali and European languages using a small 1.5B model with constrained generation.

There is also a second, unstudied factor: the language of the few-shot exemplars. Exemplars do not just show how to solve a problem; they demonstrate a reasoning style and structure the model tends to follow. The interaction between exemplar language and CoT language has not, to our knowledge, been studied. We address both gaps with the following contributions:

- We run a 2×2 factorial experiment crossing exemplar language (English/Urdu) and CoT language (English/Urdu) on MGSM Urdu and Math-500 Urdu, enabling us to measure their interaction rather than individual effects alone.

- Alongside accuracy, we measure chain length as both token count and discrete step count, testing whether chain structure changes across conditions in ways predicted by [Feng et al. \(2023\)](#) and [Amiri et al. \(2025\)](#).
- We find that exemplar language completely moderates the effect of CoT language: switching to Urdu CoT causes a 34.8 point drop with English exemplars but only 1.6 points with Urdu exemplars, and that Urdu CoT produces more tokens but not more discrete reasoning steps.

2 Background and Related Work

Chain-of-Thought as Computation. [Wei et al. \(2022\)](#) introduced CoT prompting, and [Feng et al. \(2023\)](#) provided a theoretical account: without intermediate steps, a transformer can only solve problems within a single computational pass. Adding CoT tokens gives the model more computation. [Amiri et al. \(2025\)](#) formalise this with lower bound proofs for certain problem classes. [Jin et al. \(2024\)](#) show that step count correlates with accuracy in English, and that having more steps helps even if some contain errors. We treat this as motivation for measuring chain length as a dependent variable.

Multilingual Chain-of-Thought. [Shi et al. \(2022\)](#) introduced the MGSM benchmark and showed English CoT generally outperforms native-language CoT across ten languages. [Huang et al. \(2023\)](#) found that how much English pivoting helps depends on the target language’s resource level. [Barua et al. \(2025\)](#) find similar patterns for newer reasoning models. Urdu is not covered in any of these. [Ferrao et al. \(2025\)](#) measure step-level attribution for Bengali, German, and French, finding step counts vary and correlate with accuracy, but use a single 1.5B model with constrained generation and do not manipulate exemplar language. We extend this to Urdu, use a larger model, allow natural generation, and add factorial manipulation of exemplar language.

A complementary line of work aims to improve cross-lingual reasoning rather than only characterise it, and both efforts turn on the cross-lingual alignment and structural organisation of intermediate steps that are central to our findings. [Ranaldi et al. \(2024\)](#) show that program-aided language models, which externalise intermediate

computation as executable code, transfer multi-step reasoning across languages more reliably than free-form CoT, precisely because the program format imposes an explicit, language-invariant structure on the reasoning steps. [Ranaldi and Pucci \(2025\)](#) instead use self-training on model-generated reasoning to align a model’s reasoning process across languages, narrowing the gap between English and lower-resource languages. Our results are consistent with the premise behind both: when the exemplar and reasoning languages are mismatched, the structural organisation of Urdu reasoning degrades, and we return to structure-enforcing prompting and self-training as candidate remedies in Section 5 and future work.

Urdu NLP and Reasoning. Despite over 230 million speakers, Urdu remains underrepresented in NLP benchmarks ([Shafique et al., 2026](#)). Urdu uses the Nastaliq script (right-to-left), shares characters with Arabic and Persian, and exists in complex diglossia with Hindi, meaning models trained on Hindi may transfer partially but unpredictably. [Arif et al. \(2024\)](#) compare general-purpose LLMs against fine-tuned specialists across Urdu tasks, finding that CoT prompting yields only modest gains on classification, but they do not study mathematical reasoning or chain structure. [Shafique et al. \(2026\)](#) introduce UrduBench with human-verified translations of MGSM, Math-500, CommonsenseQA, and OpenBookQA. Their leaderboard shows Gemini-class models perform best. We use UrduBench as our dataset source because of its translation quality guarantee, allowing us to attribute condition differences to prompt factors rather than translation artifacts.

3 Experimental Setup

3.1 Datasets

MGSM Urdu. We use the 250-problem test set from UrduBench ([Shafique et al., 2026](#)), which provides human-verified Urdu translations of MGSM ([Shi et al., 2022](#)). We verified all 250 answer pairs match the English originals exactly. MGSM consists of elementary math word problems requiring 2–5 reasoning steps.

Math-500 Urdu. We also evaluate on the 500-problem Math-500 Urdu subset from UrduBench, covering competition-level mathematics from the MATH benchmark ([Hendrycks et al., 2021](#)). This harder dataset tests whether condition effects on

Cond.	Exmp.	CoT	Test Q	CoT?
A	EN	EN	EN	✓
B	EN	UR	UR	✓
C	UR	EN	UR	✓
D	UR	UR	UR	✓
E	EN	None	EN	×
F	UR	None	UR	×

Table 1: Experimental conditions. All use 8 exemplars; only language of components changes.

MGSM generalise to more difficult problems.

3.2 Conditions

We use 8-shot prompting across six conditions. Conditions A–D form the 2×2 factorial; E and F are no-CoT baselines (Table 1). The exemplars are the official 8-shot MGSM demonstrations from Shi et al. (2022). For conditions requiring Urdu, we translated using Gemini 2.5 Flash and manually verified all 32 exemplar variants (8 exemplars $\times$ 4 language combinations). All prompts use temperature = 0.

3.3 Model

We evaluate Gemini 2.5 Flash throughout. This model has strong documented Urdu capability (Shafique et al., 2026), and using a single model lets us focus on prompt factors rather than confounding model differences.

3.4 Metrics

Accuracy. We extract the final numerical answer via regex and compare to the gold answer, following standard MGSM evaluation (Shi et al., 2022).

Chain length. We measure (a) token count of the reasoning portion using the cl100k tokenizer, and (b) the number of discrete reasoning steps. Token count measures verbosity, while step count is closer to what the theoretical framework means by “steps.” Our primary step measure is the count of explicit step markers (English *Step n* and its Urdu equivalent), which is language-neutral; as a complementary measure we also segment the reasoning into sentences using a script-aware splitter that treats the Urdu full stop (Unicode U+06D4) as a sentence boundary in addition to Latin punctuation. We validate both measures against manual annotation in Appendix E.

Language consistency. We run `langdetect` on reasoning text to check whether the model rea-

Cond.	MGSM		Math-500	
	Acc.	ΔA	Acc.	ΔA
A (EN, EN)	.672	0	.130	0
B (EN, UR)	.324	−.348	.086	−.044
C (UR, EN)	.484	−.188	.104	−.026
D (UR, UR)	.468	−.204	.082	−.048
Marginal means				
EN exmp	.498		.108	
UR exmp	.476		.093	
EN CoT	.578		.117	
UR CoT	.396		.084	

Table 2: Accuracy across the 2×2 factorial. ΔA shows difference from Condition A.

soned in the intended language.

4 Results

4.1 Baseline Results

On MGSM, Gemini 2.5 Flash solves 97.2% in English and 93.2% in Urdu without CoT, and adding CoT drops accuracy: English CoT (Condition A) scores 67.2%. The same pattern holds on Math-500 (no-CoT: 47.0%/41.2%; English CoT: 13.0%). This is a known phenomenon: CoT can introduce noise when the model has sufficient parametric knowledge (Sprague et al., 2025). The condition differences should thus be read as how much each configuration degrades performance.

4.2 Accuracy: 2×2 Factorial

The most striking result is the difference between Conditions A and B. Switching from English to Urdu CoT with English exemplars drops accuracy from 67.2% to 32.4%, a 34.8 percentage point fall. The same switch with Urdu exemplars (C to D) drops accuracy from 48.4% to 46.8%, only 1.6 points. The CoT language factor has a 22-fold larger effect depending on which exemplars are used.

We model accuracy with an item-level logistic regression with exemplar language, CoT language, and their interaction as predictors, using cluster-robust standard errors on problem ID to account for the same problems recurring across conditions; a mixed-effects logistic regression with a random intercept per problem yields the same conclusions (Appendix A). On MGSM this confirms a significant average main effect of CoT language ($z = -8.53, p < 0.001$) and a highly significant exemplar $\times$ CoT interaction ($z = 6.92, p < 0.001$; odds ratio = 4.0). The average main effect of ex-

Accuracy by Condition (2×2 Factorial)

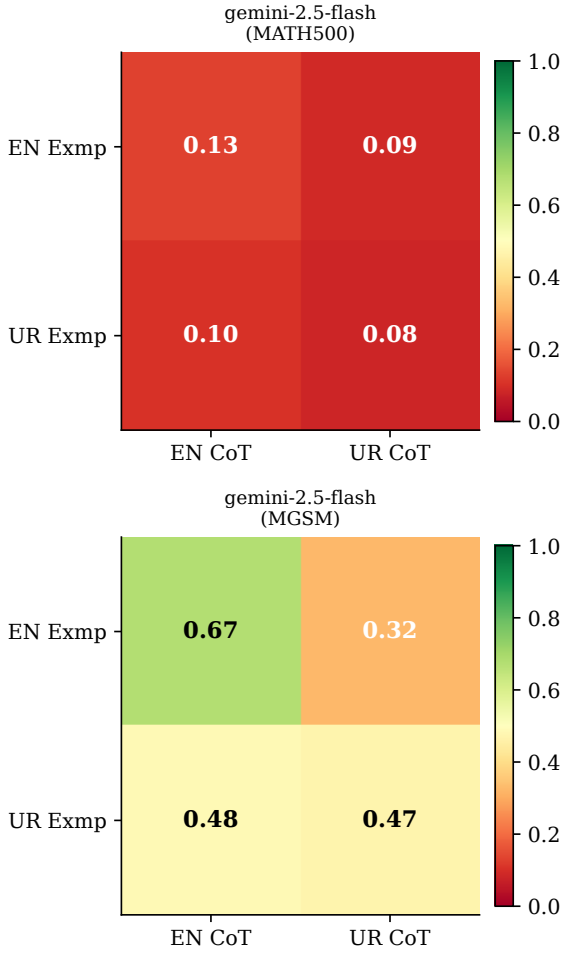


Figure 1: Accuracy heatmap across the 2×2 factorial on MGSM Urdu (bottom) and Math-500 Urdu (top). The large drop in Condition B (English exemplars, Urdu CoT) is visible on both datasets.

emplar language is not significant ($z = -0.97$, $p = 0.33$), indicating exemplar language matters only through its interaction with CoT language.

McNemar’s paired tests sharpen this: A vs. B is highly significant ($p < 0.001$), while C vs. D is not ($p = 0.694$). The B vs. D comparison is also significant ($p < 0.001$): switching from English to Urdu exemplars while keeping Urdu CoT improves accuracy by 14.4 points.

On Math-500, the same logistic regression shows the interaction is not significant ($z = 1.19$, $p = 0.24$), though CoT language remains significant ($z = -3.64$, $p < 0.001$). The narrow accuracy range (8–13%) limits statistical power. The condition ordering ($A > C > D > B$) is preserved across datasets.

Cond.	MGSM		Math-500	
	Tok.	Steps	Tok.	Steps
A (EN, EN)	91.2	3.24	50.0	1.64
B (EN, UR)	123.5	2.24	65.5	1.34
C (UR, EN)	110.5	2.79	53.6	1.74
D (UR, UR)	155.2	2.68	79.6	1.37

Table 3: Mean reasoning chain length by condition. “Steps” counts explicit step markers, validated against manual annotation (Appendix E). Urdu CoT produces more tokens but slightly fewer discrete steps, and the step reduction is concentrated in Condition B.

4.3 Chain Length Analysis

Our initial hypothesis was that Urdu CoT would produce shorter chains. Instead, Urdu CoT produces substantially more tokens (139.4 vs. 100.9 on MGSM). ANOVA confirms both CoT language ($F = 57.74$, $p < 0.001$) and exemplar language ($F = 25.30$, $p < 0.001$) have significant effects on token count, with no significant interaction ($F = 1.51$, $p = 0.22$).

However, step count does not track token count. Using explicit step markers, which agree with manual annotation at $r = 0.99$ (Appendix E), English CoT averages 3.0 discrete steps and Urdu CoT 2.5, a modest reduction. Crucially, the reduction is not uniform: it is driven almost entirely by Condition B, where 41% of responses terminate after a single step, whereas Condition D (matched Urdu exemplars) retains step structure comparable to English CoT. Urdu CoT thus adds tokens without adding proportional reasoning steps, and under exemplar–CoT mismatch it produces markedly fewer. This matters because the theoretical lower bounds of Amiri et al. (2025) are defined in terms of reasoning steps, not tokens.

4.4 Length–Accuracy Correlation

Despite Urdu CoT producing longer token sequences, longer chains still correlate positively with accuracy. Across all CoT conditions pooled, $r = 0.42$ ($p < 0.001$) on MGSM and $r = 0.39$ ($p < 0.001$) on Math-500. The correlation holds within each condition: A ($r = 0.45$), B ($r = 0.60$), C ($r = 0.35$), D ($r = 0.59$). The correlation is strongest in Urdu CoT conditions (B and D), where high variance in response length means some Urdu CoT responses happen to produce more structured output, and those tend to be correct.

This is consistent with the theoretical prediction:

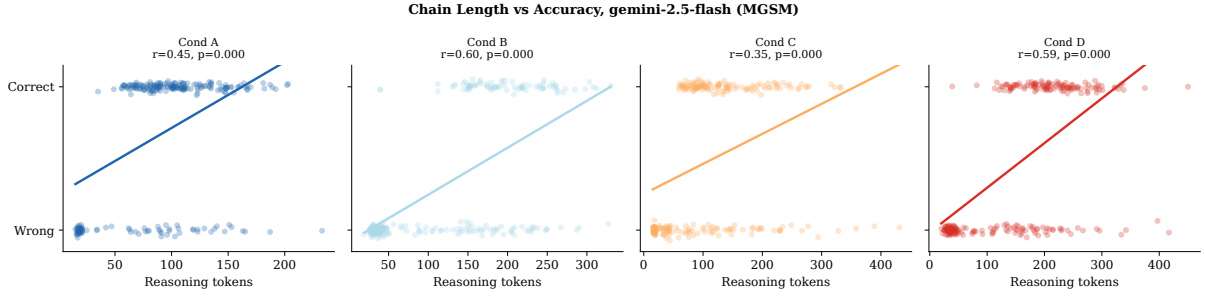


Figure 2: Chain length (tokens) vs. accuracy on MGSM by condition. Points are jittered for visibility. The positive correlation holds across all four conditions, with the strongest effect in the Urdu CoT conditions (B and D).

Condition	MGSM	Math-500
A (EN exmp, EN CoT)	0.004	0.008
B (EN exmp, UR CoT)	0.020	0.152
C (UR exmp, EN CoT)	0.276	0.058
D (UR exmp, UR CoT)	0.016	0.096

Table 4: Language switch rates by condition (proportion of responses where detected language differs from intended CoT language).

longer chains provide more opportunity for correct computation, even in Urdu. But critically, this is a within-condition correlation, not across conditions. Urdu CoT averages more tokens but fewer steps and is less accurate overall, suggesting token count alone is not a reliable proxy for computational depth across languages.

4.5 Language Consistency

Condition C (Urdu exemplars, English CoT) has the highest switch rate at 27.6% on MGSM, more than a quarter of responses drifted into Urdu despite instructions to reason in English. Condition B (English exemplars, Urdu CoT) switches only 2.0%. This asymmetry suggests exemplar language is a stronger signal than instruction language. English appears to act as a stronger attractor than Urdu, consistent with pretraining data imbalance.

Importantly, language switching does not directly explain accuracy differences. Switched and non-switched responses are within 0.6 points on MGSM (0.481 vs. 0.489). The accuracy drop in Condition B is better explained by structural degradation: the model stays in Urdu but frequently fails to complete a full sequence of discrete steps, with 41% of Condition B responses terminating after a single step.

5 Discussion

The interaction is the main result. When exemplars and CoT are in the same language (A or D), performance is relatively stable. When mismatched, consequences depend on direction: English exemplars with Urdu CoT (B) is catastrophic (32.4%), while Urdu exemplars with English CoT (C) is moderately bad (48.4%). English exemplars create a strong prior for English-style structured reasoning; when forced into Urdu, the model cannot meet that expectation. Urdu exemplars set expectations the model can actually fulfil in Urdu.

When CoT hurts. Our baselines show CoT reduces accuracy compared to no-CoT on both datasets. Sprague et al. (2025) show CoT helps mainly for math and symbolic reasoning but can hurt on other tasks; even within math, strong models on easy problems may not benefit. The mismatch between elementary exemplar difficulty and competition-level Math-500 problems likely exacerbates this. A practical takeaway: exemplar difficulty should be matched to test difficulty.

Step count as the right operationalisation. Token count and step count diverge for Urdu CoT: more tokens, but no corresponding gain in discrete steps, and a clear loss in the mismatched condition. Token count alone would suggest that Urdu CoT provides more computational depth, which the accuracy numbers contradict. The theoretical lower bounds of Amiri et al. (2025) are stated in terms of steps, not tokens. Using a step measure that we validate against manual annotation, our results support step count as the appropriate proxy for computational depth in multilingual settings, and show that raw token count can be misleading when verbosity differs across conditions.

What language switching reveals. Exemplar language appears to be a stronger signal than instruction language for controlling model output. The accuracy drop in Condition B is not primarily a language control failure (the model stays in Urdu) but a reasoning structure failure (it often fails to complete a sequence of discrete steps). These are distinct phenomena that would have been conflated without separate measurements.

Practical implications. For practitioners deploying CoT in Urdu, the key message is that exemplar choice matters more than expected. Using English exemplars with Urdu CoT, which might seem reasonable when Urdu exemplars are scarce, turns out to be the worst-performing condition. The exemplar–CoT language mismatch causes structural degradation, often reasoning that terminates prematurely, that a same-language configuration largely avoids. Where possible, exemplars and CoT should be in the same language.

6 Conclusion

We studied how exemplar language and CoT language independently and jointly affect reasoning accuracy and chain structure in Urdu. The main finding is a strong interaction: exemplar language determines whether CoT language matters at all. Switching to Urdu CoT with English exemplars causes a 34.8 point accuracy drop; the same switch with Urdu exemplars causes essentially no change.

We also found that Urdu CoT does not produce shorter chains; it produces longer ones (more tokens) without a matching increase in discrete reasoning steps, and under exemplar–CoT mismatch it produces markedly fewer, often incomplete, steps. This points to a validated step count, not token count, as the appropriate proxy for computational depth as defined by Amiri et al. (2025). Future work should examine whether this finding holds for other low-resource languages and task types, and whether structure-enforcing approaches such as program-aided prompting (Ranaldi et al., 2024) or self-training on structured reasoning traces (Ranaldi and Pucci, 2025) can restore step-organised Urdu output.

Limitations

We evaluate a single model (Gemini 2.5 Flash), so we cannot determine how general these findings are across architectures or model sizes. We only examine mathematical reasoning tasks; the

exemplar–CoT interaction may look different for other task types such as commonsense or logical reasoning. The length–accuracy correlation is consistent with the theoretical framework but does not establish causation: longer chains may simply reflect higher model confidence on easier problems. Our discrete-step measure relies on explicit step markers together with script-aware sentence segmentation; while these agree closely with manual annotation on a stratified sample of 48 responses (Appendix E), the manual annotation was performed by the author, and a larger, independently annotated set would further strengthen the validation. Responses that reason without explicit markers are captured only by the sentence-based measure, which may still undercount steps in continuous prose. The `langdetect` library is imperfect for Urdu due to shared Nastaliq script features with Hindi and Persian, and some language classifications may be incorrect. Finally, our exemplar translations were produced by Gemini 2.5 Flash and manually verified, but we cannot rule out subtle translation effects.

References

- Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *Proceedings of the 42nd International Conference on Machine Learning*.
- Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024. Generalists vs. specialists: Evaluating large language models for Urdu. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7263–7280.
- Josh Barua, Seun Eisepe, Kayo Yin, and Alane Suhr. 2025. Long chain-of-thought reasoning across languages. *arXiv preprint arXiv:2508.14828*.
- Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 36.
- Jeremias Ferrao, Ezgi Basar, Khondoker Ittehadul Islam, and Mahrokh Hassani. 2025. What really counts? examining step and token level attribution in multilingual CoT reasoning. *arXiv preprint arXiv:2511.15886*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*.

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

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582.

Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. 2024. Empowering multi-step reasoning across languages via program-aided language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12171–12187.

Muhammad Ali Shafique, Areej Mehboob, Layba Fiaz, Muhammad Usman Qadeer, and Hamza Farooq. 2026. UrduBench: An Urdu reasoning benchmark using contextually ensembled translations with human-in-the-loop. *arXiv preprint arXiv:2601.21000*.

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. 2022. [Language models are multilingual chain-of-thought reasoners](#).

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

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

A Statistical Analysis Tables

Table 5 reports an item-level logistic regression for accuracy with effect-coded (sum-to-zero) pre-

dictors, so the exemplar- and CoT-language coefficients are average main effects. We use cluster-robust standard errors on problem ID to account for the same problems recurring across conditions. As a robustness check, a mixed-effects logistic regression with a random intercept per problem produces the same signs and significance pattern (e.g., on MGSM the exemplar $\times$ CoT interaction has posterior mean 2.77 on the log-odds scale, clearly separated from zero).

Dataset	Factor	z	p
MGSM	Exemplar lang.	−0.97	0.331
	CoT lang.	−8.53	<0.001
	Exemplar $\times$ CoT	6.92	<0.001
Math-500	Exemplar lang.	−1.46	0.145
	CoT lang.	−3.64	<0.001
	Exemplar $\times$ CoT	1.19	0.235

Table 5: Item-level logistic regression for accuracy (effect-coded predictors; cluster-robust standard errors on problem ID). The MGSM interaction corresponds to an odds ratio of 4.0 under treatment coding.

We report a two-way ANOVA for reasoning-token length, a continuous outcome (Table 6).

Dataset	Factor	F	p
MGSM	Exemplar lang.	25.30	<0.001
	CoT lang.	57.74	<0.001
	Exemplar $\times$ CoT	1.51	0.220
Math-500	Exemplar lang.	8.05	0.005
	CoT lang.	43.70	<0.001
	Exemplar $\times$ CoT	2.81	0.094

Table 6: Two-way ANOVA for reasoning-token length (continuous outcome).

B McNemar’s Test Results

Pair	Varies	Acc ₁	Acc ₂	p
A vs B	CoT (EN exmp)	.672	.324	<.001
C vs D	CoT (UR exmp)	.484	.468	.694
A vs C	Exmp (EN CoT)	.672	.484	<.001
B vs D	Exmp (UR CoT)	.324	.468	<.001

Table 7: McNemar’s tests on MGSM ($n = 250$).

C No-CoT Baselines

Dataset	E (No CoT, EN)	F (No CoT, UR)	A (EN CoT)	D (UR CoT)
MGSM	0.972	0.932	0.672	0.468
Math-500	0.470	0.412	0.130	0.082

Table 8: No-CoT baselines vs. best/worst CoT conditions.

D Example Responses

Condition A (correct, structured). For the standard “Janet’s ducks” problem, the model produces:

```
Step 1: Janet's ducks lay 16
eggs per day.
Step 2: She eats 3 for
breakfast: 16 - 3 = 13.
Step 3: She bakes muffins with
4: 13 - 4 = 9.
Step 4: She sells 9 eggs at $2
each: 9 x 2 = $18.
The answer is: 18
```

Condition B (bimodal; frequent truncation).

Condition B does not uniformly abandon step structure. Many responses still emit explicit numbered steps in Urdu (Step 1, Step 2, ...): 44% of Condition B responses contain three or more explicit steps. However, 41% terminate after a single, often incomplete, step. A representative truncated response (translated) reads: “Step 1: Determine the number of blue ties. *” and then stops before completing the computation. This premature termination, rather than a wholesale switch to unstructured prose, is what drives the low mean step count and the accuracy drop in Condition B.

E Step-Count Validation

To validate our discrete-step measures, we drew a stratified random sample of 48 MGSM CoT responses (12 per condition, A–D) and manually annotated the number of discrete reasoning steps in each response. We then compared three automatic measures against these manual counts: (i) a sentence count using a Latin-punctuation splitter, (ii) a script-aware sentence count that also treats the Urdu full stop (U+06D4) as a boundary, and (iii) explicit step markers. Table 9 reports Pearson correlation and mean absolute error (MAE) against the manual counts, both overall and restricted to the Urdu-CoT conditions (B and D) where script effects are strongest.

Measure	All ($n=48$)		UR CoT ($n=24$)	
	r	MAE	r	MAE
Latin-only sentences	0.57	1.17	0.18	1.21
Script-aware sentences	0.77	0.92	0.84	0.50
Explicit step markers	0.99	0.04	0.99	0.04

Table 9: Agreement of automatic step measures with manual annotation ($n = 48$; 12 per condition). A Latin-punctuation sentence splitter undercounts Urdu steps by ~ 1.2 on average and correlates only weakly with manual counts within Urdu-CoT conditions, because Urdu sentences terminate with a different full stop. Explicit step markers track manual counts almost exactly, so we adopt them as the primary step measure and report script-aware sentence counts as a complement.

This undercounting is systematic rather than random: because Urdu sentences terminate with the Urdu full stop (U+06D4) rather than a Latin period, multi-step Urdu chains are read as a single “sentence.” The script-aware splitter removes most of this bias, and explicit step markers, which count language-neutral `Step/step` tokens, agree with human annotation almost perfectly.

F Additional Figures

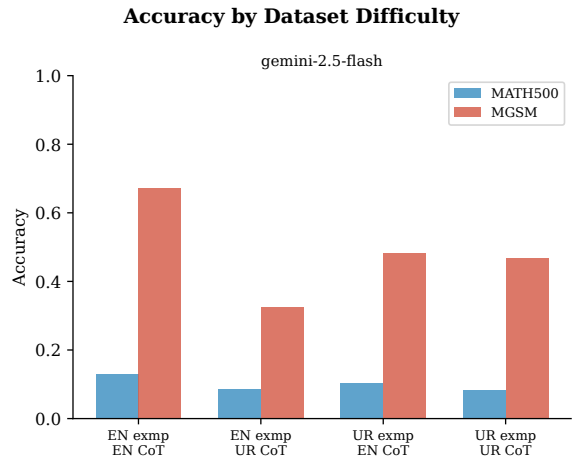


Figure 3: Accuracy by condition on MGSM vs. Math-500. The relative ordering of conditions is preserved but the absolute differences compress on the harder dataset.

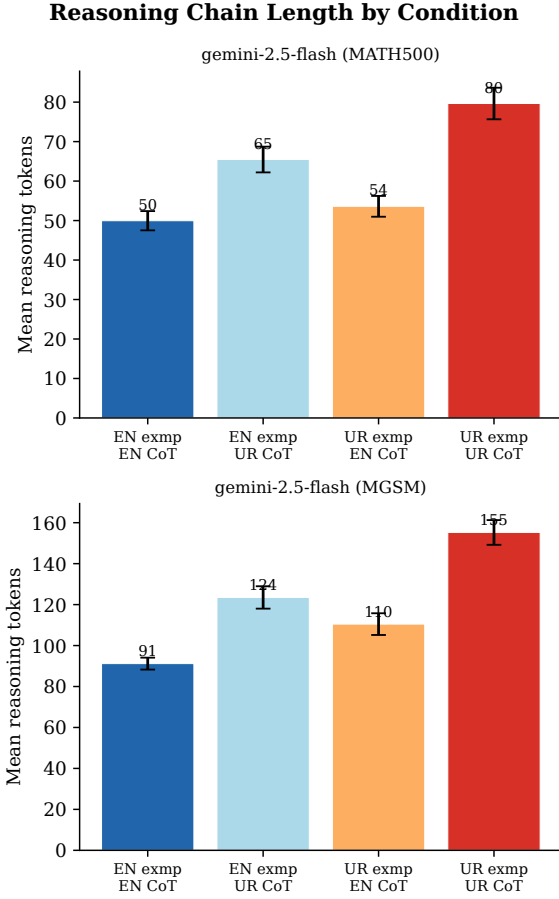


Figure 4: Mean reasoning chain length (tokens $\pm$ SE) by condition. Urdu CoT conditions (B, D) produce more tokens than English CoT conditions (A, C), without a corresponding increase in discrete reasoning steps (Appendix E).

G Prompt Structure

All six conditions share the same structure: a system instruction, 8 few-shot exemplar blocks, and the test question with answer prefix. The English CoT instruction reads: “Solve the following math problem step by step. Show your reasoning clearly, then state the final answer as a number.” The no-CoT instruction reads: “Answer the following math problem. Give only the final numerical answer.” Urdu instructions are direct translations.

Cond.	Instr.	Exmp Q	Exmp CoT	Test Q
A	EN	EN	EN	EN
B	UR	EN	UR	UR
C	EN	UR	EN	UR
D	UR	UR	UR	UR
E	EN	EN	None	EN
F	UR	UR	None	UR

Table 10: Language of each prompt component.