Evaluating Cross-lingual Knowledge Consistency in Code-Mixed vis-a-vis Indian Languages using IndicKLAR

Anonymous ACL submission

Abstract

Large language models recall knowledge reliably in English but often fail on the same query posed in a lower-resourced language—a crosslingual consistency gap that remains underexplored for Indian languages and their code-mixed counterparts. To study this gap, we introduce **INDIKLAR**, an Indic extension of the KLAR-CLC benchmark covering 18 of the 22 scheduled Indian languages and pairing them with code-mixed variants for 11 widely used language pairs, with native-speaker verification of both monolingual and code-mixed variants for these 11 settings. This three-way alignment offers a unique opportunity to examine how knowledge recall consistency varies across the spectrum of English, code-mixed, and native Indian language inputs. Evaluating across nine open-weight models, we find that the native-language accuracy gap to English can reach ∼0.50, while code-mixed inputs close most of it—bringing performance within ∼0.05 of English without any model-level intervention. Motivated by this, we evaluate several prompting strategies that vary in how language conversion is exposed, including a two-stage translate-then-answer setup, a one-stage joint translation-and-answer prompt, and ***Translate-in-Thought (TinT)***—a single-step strategy in which the model converts the input internally and emits only the final answer. Across the performance trajectory native → code-mixed → English, we identify a consistent *flip point*—the boundary between incorrect and correct prediction—that lies between the native and code-mixed settings. Interestingly, this holds whether the trajectory is induced by the input surface form or by the model's internal conversion process.

1 Introduction

Large language models (LLMs) are increasingly deployed in multilingual contexts, yet their capabilities remain unevenly distributed across languages (Fu and Liu, 2025). Performance is consistently highest for high-resource languages like English, while significantly degrading for low-resource languages—a disparity that poses a critical barrier for equitable AI deployment in linguistically diverse regions (Wang et al., 2025a). To build truly equitable multilingual systems, we must evaluate and improve *crosslingual knowledge consistency*: whether models maintain consistent knowledge recall across language variants of the same query.

This work addresses crosslingual consistency in the Indian language ecosystem, a setting of significant practical importance affecting over 1.4 billion people. India recognizes 22 scheduled languages[1] spanning multiple language families, yet most LLMs remain substantially undertrained on them. Moreover, code-mixing between English and Indian languages is common in everyday digital communication—a phenomenon largely absent from existing crosslingual benchmarks, which overwhelmingly focus on clean monolingual inputs.

To fill this gap, we introduce **INDIKLAR**, an Indic extension of KLAR-CLC (Wang et al., 2025b)—a crosslingual factual consistency benchmark built around structured knowledge-base triples—covering 18 of the 22 scheduled Indian languages alongside 11 code-mixed variants constructed with English. INDIKLAR is the first benchmark to pair monolingual Indic factual queries with code-mixed counterparts under native-speaker verification of both, at this scale. This three-way alignment enables aligned evaluation across English, native Indian languages, and naturally motivated code-mixed inputs. The code-mixed variants are constructed by interleaving English content words into the grammatical structure of each Indian language, reflecting real-world mul-

[1] https://en.wikipedia.org/wiki/Languages_with_official_recognition_in_India

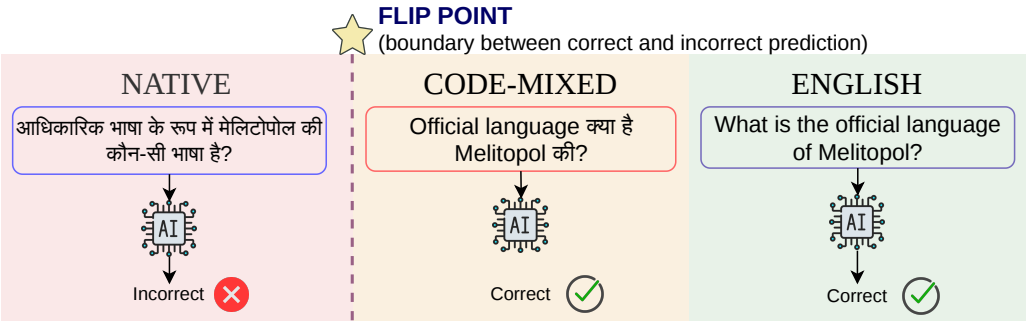


Figure 1: Illustration of the *flip point*: the boundary between incorrect and correct prediction, lies between the native Indian languages and code-mixed settings.

tilingual communication patterns. Both the monolingual and code-mixed variants undergo human quality verification by native or near-native speakers across all 11 language settings.

Our evaluation on INDIKLAR reveals a pronounced consistency gap between native and English inputs—one that, surprisingly, code-mixed variants largely close without any model-level intervention.

Drawing inspiration from bilingual cognition—where speakers responding in a non-native language often process internally in their dominant one (Grosjean, 2008; Kroll and Stewart, 1994; Green and Abutalebi, 2013)—we ask whether LLMs possess a latent translation capability that can be elicited through prompting alone. This motivates **Translate-in-Thought (TinT)**, a single-step strategy in which the model converts the input internally and emits only the final answer. Alongside TinT, we evaluate two-step and one-step joint prompting strategies, systematically varying the visibility of the intermediate conversion.

We summarize our contributions as follows:

- We introduce INDIKLAR, an Indic extension of KLAR-CLC covering 18 of the 22 scheduled Indian languages and code-mixed variants for 11 widely used language pairs—human-verified across both monolingual and code-mixed forms—and conduct a systematic consistency analysis showing that code-mixed inputs substantially reduce native-language performance gaps.

- We introduce Translate-in-Thought (TinT) prompting, a single-step implicit strategy that elicits internal translation without explicit intermediate outputs, and benchmark it against baseline and other prompting techniques.

- Our study reveals that the *flip point*—the boundary between incorrect and correct prediction—consistently falls between the native low-resource and code-mixed settings, whether viewed as a property of the input surface form or the model's internal thought process.

2 Related Work

Crosslingual Consistency Benchmarks. Existing benchmarks for crosslingual factual consistency predominantly focus on high-resource and globally distributed language pairs, leaving the Indian language ecosystem underrepresented (Qi et al., 2023; Wang et al., 2025b). These benchmarks cover few Indian languages and rarely include code-mixed variants that reflect real-world communication patterns. While Han et al. (2025) and Liu et al. (2025) include limited evaluation on code-switched inputs, their code-switching is constructed systematically via direct placeholder substitution, which does not reflect naturalistic code-switching observed in real-world multilingual communication—and may partly explain why Han et al. (2025) report inconsistent gains on code-switched variants compared to monolingual baselines. More broadly, existing evaluation frameworks such as Gupta et al. (2025) rely on translating model outputs back into English via large external systems such as NLLB-54B before comparison against gold labels—a setup that can obscure whether the low-resource output was itself faithful, and makes the evaluation pipeline comparatively expensive. INDIKLAR addresses these gaps by providing native-language evaluation across 18 of the 22 scheduled Indian languages with native-speaker verification of monolingual and code-mixed variants for 11 of them, without dependence on external translation systems.

Prompting for Multilingual Reasoning. On the methods side, prior work has largely approached low-resource multilingual performance through training-time interventions. Ai et al. (2025) find that code-switching training and cross-lingual word alignment objectives yield the most consistent improvements across languages. However, they do not examine how inference-time prompting strategies affect model behavior in code-mixed settings. Kim et al. (2025) argue that activating the right internal mechanisms—including better neuron-level utilization and internal translation behavior—can substantially improve how existing multilingual knowledge is used. This suggests that prompting strategies alone may unlock stronger crosslingual consistency without retraining. The approach of Wang et al. (2025b) requires extensive hyperparameter search across language-specific configurations, making it costly to adapt to new models and languages at scale. They further argue that translation-based methods are ineffective for crosslingual factual prediction. Our results directly challenge this claim, showing that translation used as an intermediate substrate substantially improves over direct prompting.

These data- and methods-side gaps jointly motivate our work. INDIKLAR provides broad Indic coverage with human-verified code-mixed variants. On top of this, our prompting study—including Translate-in-Thought (TinT)—evaluates crosslingual consistency as a native model capability rather than an artifact of external linguistic support.

3 IndiKLAR Benchmark

We extend KLAR-CLC (Wang et al., 2025b) to 18 of the 22 scheduled Indian languages, creating language-aligned instances for consistent evaluation across multiple language settings.

Monolingual Indian Language Variants. We translate English KLAR-CLC into 18 Indian languages using the Google Translator API[2], yielding 18 Indian-language variants aligned with the original English source : Assamese (as), Bengali (bn), Dogri (doi), Gujarati (gu), Hindi (hi), Kannada (kn), Konkani (gom), Maithili (mai), Malayalam (ml), Marathi (mr), Nepali (ne), Odia (or), Punjabi (pa), Sanskrit (sa), Sindhi (sd), Tamil (ta), Telugu (te), and Urdu (ur). The languages with their ISO codes are presented in Table 4.

Code-Mixed Variants. For 11 out of the 18 Indian languages—those for which we had linguist coverage or native speaker (Assamese, Bengali, Gujarati, Hindi, Malayalam, Marathi, Odia, Punjabi, Sanskrit, Tamil, and Telugu)—we construct code-mixed variants generated using GPT APIs[3]. Both the monolingual and code-mixed variants for these 11 languages were manually verified by native speakers for fluency, and naturalness of English–Indic mixing; annotator demographics are presented in §7.

Dataset Statistics. INDIKLAR provides three-way parallel variants (monolingual native Indian, code-mixed, and English) for 11 languages, and two-way parallel variants (monolingual native Indian and English) for the remaining 7. Following KLAR-CLC (Wang et al., 2025b), each language setting contains 2,619 instances spanning 20 fact categories. Number of samples, average query lengths, and average answer lengths are reported in Table 5 in the Appendix.

Alignment and Release. The three-way alignment enables experimentation across multiple dimensions simultaneously. It lets us compare model behavior not just between high- and low-resource settings, but also along the intermediate axis of code-mixed input, all while keeping the underlying semantic content fixed across variants. We will publicly release INDIKLAR to facilitate future research.

4 Experiments

4.1 Baseline and Trajectory

We evaluate model robustness across three input settings: (i) **English (EN)**—the high-resource upper bound, (ii) **Native (NA)**—the query in original script, and (iii) **Code-Mixed (CM)**—the query in code-mixed format with Roman script. This comparison is designed to test whether code-mixing enables models to leverage stronger English reasoning while preserving low-resource semantics.

4.2 Prompting Strategies

Beyond direct prompting, we investigate several conversion-based strategies designed to bridge the crosslingual gap, categorized into three groups.

2-Step Strategies: These decouple the transformation and answering tasks into two separate inference calls. *2Step-CM* first converts the query

[2] https://cloud.google.com/translate?hl=en

[3] https://chatgpt.com/

Strategy	Transformation	Procedure
Base (EN/-NA/CM)	None	Direct prompting on the source query in the given language.
2Step-CM	Code-mix conversion	Convert to code-mixed form, then answer in a second call.
2Step-EN	English translation	Translate to English, then answer in a second call.
2Step-Translit	Roman transliteration	Transliterate to Roman script without translation, then answer in a second call.
1Step-CM+Ans	Code-mix conversion	Emit conversion and answer in a single pass.
1Step-EN+Ans	English translation	Emit translation and answer in a single pass.
TinT-CM	Latent code-mixed reasoning	Reason internally in code-mix; output only the answer.
TinT-EN	Latent English reasoning	Reason internally in English; output only the answer.

Table 1: Prompting strategies evaluated in our experiments. All start from the source query.

into a code-mixed form, while *2Step-EN* translates it fully into English, after which the final answer is generated from the reformulated input. *2Step-Translit* transliterates the query into Roman script without translation, isolating the effect of script conversion alone.

1-Step Strategies: These request both the transformation and the final answer in a single prompt. *1Step-CM+Ans* generates a code-mixed conversion alongside the answer, and *1Step-EN+Ans* does the same with an English translation.

Translate-in-Thought (TinT): The TinT strategy instructs the model, via a single user-prompt directive, to internally reason through an English (*TinT-EN*) or code-mixed (*TinT-CM*) representation and emit only the final answer—unlike explicit chain-of-thought, no intermediate text is generated. This tests the model's ability to utilize latent multilingual reasoning while maintaining inference efficiency.

Detailed prompts are presented in Appendix D.

4.3 Experimental Setup

Models: We evaluate open-weight language models from multiple families: *Llama 3.1*, *Llama 3.2*, *Gemma 3*, and *Qwen 2.5*, spanning roughly 1B–14B parameters. Per-model details are provided in Appendix Table 3.

Metrics: We report two primary metrics: (i) Accuracy—the percentage of correct predictions, and (ii) Cross-Lingual Consistency (CLC)—a measure of prediction overlap across language pairs. These metrics are defined in Appendix A.1.

Reproducibility: All models are evaluated with greedy decoding (temperature=0, top-p=1.0, `max_new_tokens`=10) using each model's official chat template, with 3-shot in-context examples drawn from the English source split. Experiments use a single fixed seed (12345) and run on single NVIDIA A100 GPU (80GB).

5 Results

We investigate how model performance varies across three input conditions—native script, code-mixed (CM), and English—across nine models spanning three families, and organize the discussion by the question each subsection answers: the English–Native gap (§5.1), prompting strategies (§5.2), model-size scaling (§5.3), language-family and category effects (§5.4), pairwise CLC (§5.5), romanization and comparison with prior work (§5.6), and generalization and the location of *flip point* (§5.7).

5.1 The English–Native Gap

Figure 2 reveals two consistent patterns. First, the Native tier exhibits severe degradation compared to English, particularly for smaller models, where accuracy scores approach 0.10 on Assamese and Telugu, suggesting script-level tokenization failure rather than a lack of language understanding. Second, CM scores consistently approach English-level performance across all languages and model scales, with the per-tier mean gap remaining within $\sim$0.05, compared to a Native–English gap of up to $\sim$0.50. This indicates that code-mixing effectively bypasses the low-resource penalty at inference time, without any model-level intervention.

5.2 Prompting Strategies

Figure 3 consolidates the performance scores across 11 language pairs.

Across families, selected translation-based and code-mixed strategies reduce the gap to English, often yielding stronger performance than direct low-resource prompting. Our proposed **TinT** prompting shows statistically significant gains over the baseline (see Appendix A.5) while also offering the practical benefit of lower inference time compared to other prompting strategies (Appendix A.2).

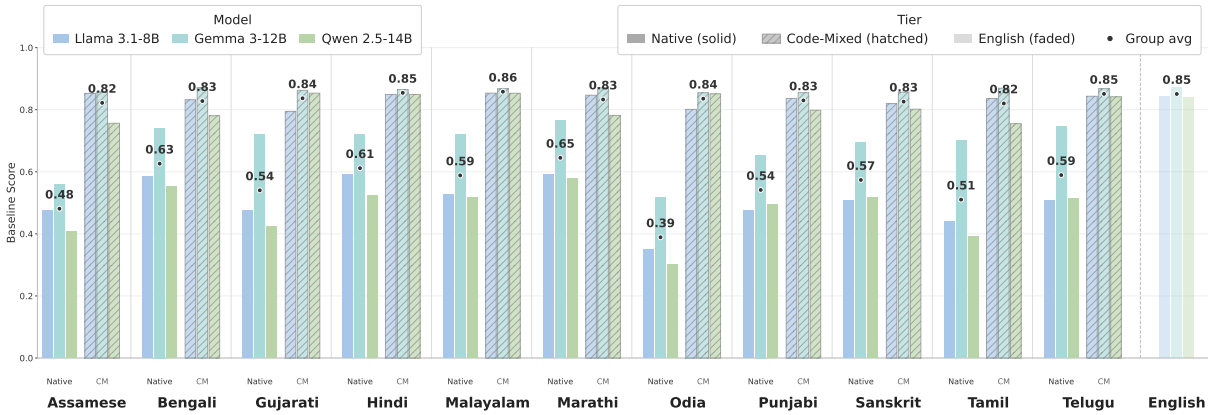


Figure 2: Baseline accuracy scores across Native (solid), Code-Mixed (hatched), and English (faded) input tiers for eleven Indic languages and English (En) using LLaMA-3.1-8B, Gemma-3-12B, and Qwen-2.5-14B. Black dots denote the per-tier group average across models.

5.3 Model-Size Scaling

Figure 5 reveals a consistent scaling trend across all three model families: as model size increases, both accuracy and CLC improve monotonically for TinT, with TinT uniformly outperforming the baseline at every scale—mean accuracy rises from 0.26 at 1B to 0.71 at 8B for Llama, and from 0.39 to 0.85 for Gemma over the same range. The gap between TinT and the baseline widens with model size, suggesting that larger models possess stronger latent translation ability. Gemma leads across all sizes, followed by Qwen and Llama. At the lower end, prompting strategies occasionally underperform the baseline for very small models (<1.5B), indicating a minimum capacity threshold below which reformulation is difficult. This degradation is not universal: while `Llama-3.2-1B` and `Qwen2.5-1.5B` show drops under several strategies, `Gemma-3-1B` consistently improves across all strategies and languages, possibly owing to stronger multilingual representations induced by its pretraining data composition. More details are provided in Appendix A.4.

5.4 Language Family and Category Effects

Language Groups. We group the evaluated languages into two families: Indo-Aryan (`as`, `bn`, `doi`, `gu`, `hi`, `gom`, `mai`, `mr`, `ne`, `or`, `pa`, `sa`, `sd`, `ur`) and Dravidian (`kn`, `ml`, `ta`, `te`). TinT achieves stronger gains for Indo-Aryan languages (avg. +18% accuracy, +14% CLC) compared to Dravidian languages (+14% accuracy, +13% CLC).

Category Analysis. We analyze TinT prompting performance across 20 fact categories per language using radar charts to compare baseline and TinT accuracy. Figure 6 reveals: (i) TinT consistently improves across most categories for all languages. (ii) Hindi and Bengali show uniform improvement, indicating TinT generalizes well to better-resourced languages. (iii) Assamese shows uneven gains (15-20% for some categories, minimal for others), suggesting inconsistent coverage for low-resource facts. (iv) Odia shows minimal improvement, indicating implicit translation cannot overcome fundamental representation gaps. (v) Semantic and cultural fact categories remain challenging across all languages, highlighting persistent limitations.

5.5 Pairwise Cross-Lingual Consistency

To understand whether models maintain consistent predictions across all language pairs, we compute pairwise cross-lingual consistency (CLC) scores comparing baseline and TinT-CM predictions for all 18 languages using Llama-3.1-8B-Instruct. Figure 4 reveals three key patterns. (i) Baseline CLC values span a wide range with high variance (0.4–0.8), whereas (ii) TinT-CM consistently improved scores, converging CLC scores to a higher range (0.5–0.9), suggesting that implicit translation enforces higher range cross-lingual consistency. (iii) Notably, Odia emerges as an outlier with the lowest CLC scores in both baseline and TinT-CM settings (0.38–0.60), likely due to limited pretraining data and script complexity, suggesting that implicit translation alone cannot overcome fundamental representational gaps for severely under-resourced languages.

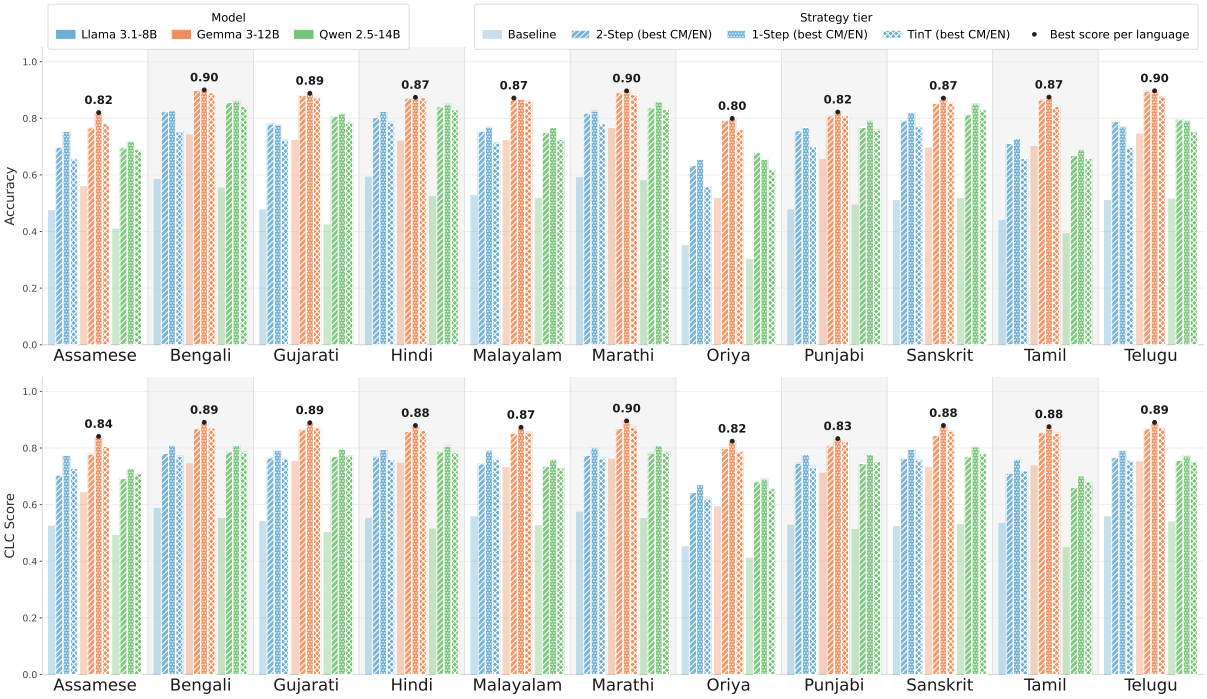


Figure 3: Performance of all prompting strategies across eleven Indian languages on two evaluation metrics: accuracy (top row) and CLC score (bottom row). The numerical scores for each model are reported in Appendix C.

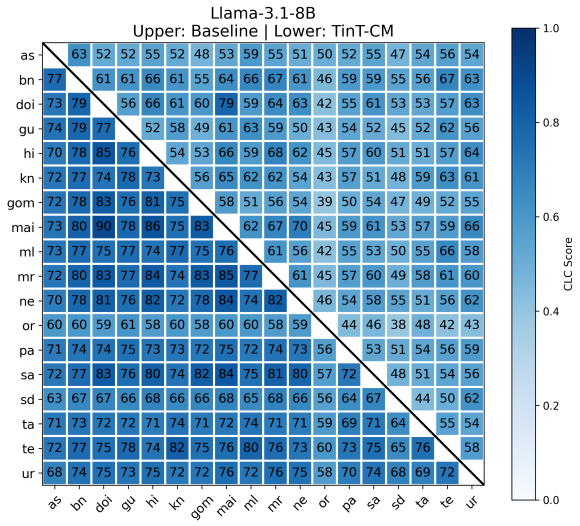


Figure 4: Pairwise cross-lingual consistency (CLC) comparison for **Llama-3.1-8B-Instruct**. The upper triangular region shows the baseline CLC scores, while the lower triangular region corresponds to **TinT-CM**.

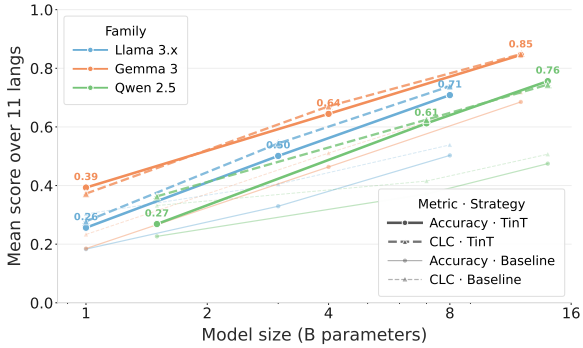
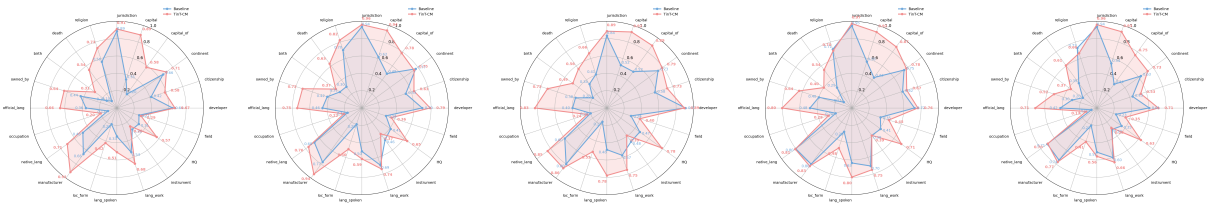


Figure 5: Mean Accuracy and CLC of TinT and Baseline strategies across 11 Indic languages.

5.6 Romanization Ablation and Literature Comparison

Ablation of code-mixing vs. romanization. To isolate whether performance gains from code-mixed conversion arise from lexical English anchors or from script uniformity alone, we compare three variants: baseline (native script), romanized (full transliteration to Latin script), and code-mixed (English-Indic mixing). Figure 7 shows that code-mixed variants substantially outperform both native-script (low-res.) and romanized baselines across languages, with improvements of 10-20% accuracy. Notably, romanized variants provide sub-optimal gains over native-script inputs, indicating that the benefits of code-mixing stem from meaningful bilingual lexical cues rather than script conversion alone.

Comparison with prior work. Table 2 shows that our approach achieves significantly higher accuracy compared to previous crosslingual consistency evaluation methods.



(a) Assamese (b) Bengali (c) Hindi (d) Marathi (e) Telugu

Figure 6: Accuracy of different category of facts across languages. Here, we present scores of only five representative languages, rest are shown in Figure 10 in Appendix.

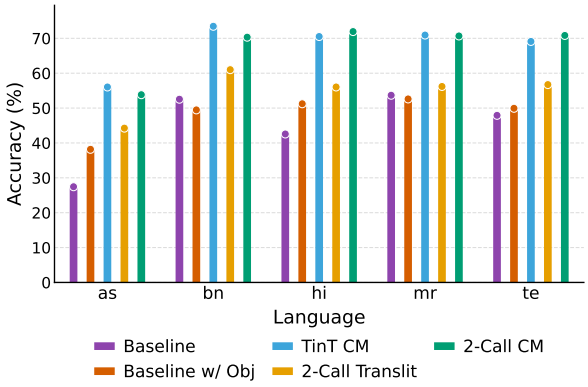


Figure 7: Ablation isolating the effect of code-mixing from romanization.

	AS		BN		HI		MR		TE	
Method	**ACC**	**CLC**	**ACC**	**CLC**	**ACC**	**CLC**	**ACC**	**CLC**	**ACC**	**CLC**
Liu et al. (2025)	0.533	0.614	0.680	0.692	0.589	0.632	0.649	0.691	0.635	0.669
CoT	0.496	0.574	0.641	0.655	0.656	0.643	0.664	0.649	0.523	0.592
Ours	0.656	0.741	0.753	0.796	0.786	0.783	0.781	0.791	0.697	0.764

Table 2: Comparison with prior work, including Chain-of-Thought (CoT) (Wei et al., 2022) prompting, on five languages.

5.7 Generalization and Flip Point

Generalization. TinT generalizes beyond INDIKLAR: evaluated on a contextually mediated knowledge recall benchmark (Liu et al., 2026) translated into five representative Indian languages, TinT consistently improves over the baseline across all languages and model families (see Appendix B).

Locating the Flip Point. Our experiments reveal a consistent pattern across languages and model families. Gold code-mixed inputs (explicit evaluation) bring performance close to English-level accuracy, while TinT prompting (implicit evaluation) independently recovers a significant portion of the same gap—suggesting that predictions are not fundamentally derailed in code-mixed settings regardless of the pathway. This leads us to a central finding: the *flip point*, marked by a sharp jump from incorrect to correct prediction, lies not between code-mixed and English, but between the native and code-mixed settings in the performance trajectory of native → code-mixed → English.

6 Analysis

6.1 Interpretability Case Study

We focus on a single representative model (Llama-3.1-8B-Instruct); the trends are averaged across all 18 Indian languages and should be read as illustrative rather than exhaustive.

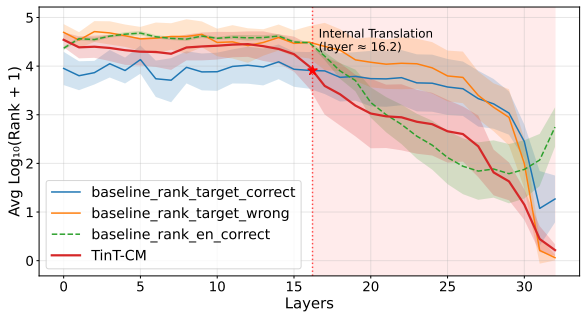


Figure 8: Layer-wise rank of the gold answer token among the model's output distribution for **Base** prompting and the best **TinT** variant across layers of Llama-3.1-8B-Instruct. Shaded bands show the standard deviation across languages.

At each layer, we compute the rank of the gold answer token within the model's predicted distribution: a rank of 1 means the model assigns it the highest probability, while higher ranks indicate it is deprioritised relative to other tokens (see Appendix A.1 for details). Figure 8 tracks four such trajectories over samples the baseline answered *incorrectly*: under baseline, the correct (`target_correct`) and finally-generated incorrect (`target_wrong`) target tokens and the English equivalent of the correct answer (`en_correct`); and the correct target token under `TinT-CM` on the same indices. The TinT-CM and baseline curves remain largely indistinguishable through the first

$\sim$16 layers. After this, TinT-CM separates: the gold answer's rank drops steeply toward zero in later layers, while the baseline's stays elevated. This gap is concentrated in the deeper layers, consistent with late-layer cross-lingual reorganization—though not by itself direct evidence of it.

6.2 Error Analysis

To understand why low-resource recall fails and how conversion-based strategies help, we qualitatively compare prompting behaviors across four languages (representative examples in Figure 9 in Appendix). It shows why surface form and conversion matter for knowledge recall.

English queries consistently yield correct answers, while **Native**-script inputs frequently fail. **2Step-Translit** reveals that script conversion alone is insufficient: although transliteration preserves pronunciation, phonetic mapping into Roman characters often distorts entity spellings.

1Step-CM+Ans and **TinT-CM**, by contrast, converts the samples to respective code-mixed variant. This partial lexical grounding preserves entity names more faithfully than transliteration, improving both semantic alignment and answer accuracy. The key distinction is that code-mixing introduces *meaningful* English anchors, whereas transliteration introduces only orthographic approximations.

Here, we present one qualitative example where TinT succeeded,

Query: रोनाल्डो किस देश का नागरिक है?
(*Which country is Ronaldo a citizen of?*)

Baseline: पुर्तगाल (Portugal) ✗

1Step+Ans, TinT: ब्राज़िल (Brazil) ✓

intended to illustrate rather than quantify the mechanism: when asked "Which country is Ronaldo a citizen of?" in Hindi, baseline prompting incorrectly predicts Portugal, ignoring the candidate list and relying instead on memorized associations with Cristiano Ronaldo. Pure transliteration (romanizing to Latin script) constrains the model to the candidate set but still predicts incorrectly. However, *1Step+Ans* converts the Hindi word of 'citizen' to English while preserving grammatical structure, and yields the correct prediction. Notably, *TinT* prompting achieves the same correct result with latent translation. We do not claim this single example is representative of the dominant failure mode; the per-category numbers in §5.4 provide the aggregate picture. Rest of the analysis section is presented in Appendix A.7.

7 Conclusion

We introduced INDIKLAR, an Indic extension of KLAR-CLC covering 18 of the 22 scheduled Indian languages and code-mixed variants for 11 widely used language pairs—human-verified across both monolingual and code-mixed forms—enabling systematic evaluation of crosslingual knowledge consistency across a linguistically diverse and practically significant setting. Our evaluation reveals a pronounced consistency gap between English and native Indian language settings, whereas code-mixed variants consistently narrow this gap.

To further reduce this gap without requiring an explicit translation step, we introduced *Translate-in-Thought (TinT)* prompting, a single-step implicit strategy that instructs models to perform language conversion internally without explicit intermediate outputs. TinT shows statistically significant gains over the baseline across languages and model families, while also offering lower inference time compared to other prompting strategies. We further show that TinT generalizes beyond INDIKLAR to contextually mediated knowledge recall, suggesting that implicit translation prompting is broadly effective for crosslingual factual knowledge recall. The effectiveness of TinT improves with model size, which we attribute to larger models possessing stronger latent translation ability. Our interpretability case study (§6.1) shows that the layer-wise rank gap between TinT and the baseline is concentrated in the deeper layers, consistent with late-layer cross-lingual reorganization. Taken together, our findings suggest that the *flip point*—the boundary between incorrect and correct prediction—lies between the native and code-mixed settings in the performance trajectory of low-resource $\rightarrow$ code-mixed $\rightarrow$ English, whether viewed as a property of the input surface form or the model's internal reasoning process. Future work includes expanding coverage to the remaining scheduled Indian languages and additional task formats (reasoning, generation), and connecting TinT's behavioral gains to mechanistic evidence in the model's internal representations.

Limitations

While INDIKLAR covers 18 of the 22 scheduled Indian languages, the monolingual translations and

code-mixed variants are synthetically generated, which may not fully capture the natural variation and dialectal diversity present in real-world multilingual communication. Although both monolingual and code-mixed variants were human-verified for 11 language settings, the monolingual variants of the remaining 7 languages were not subject to manual review. The four scheduled Indian languages not covered by INDIKLAR—Bodo, Kashmiri, Manipuri/Meitei, Santali— were excluded primarily due to limited translation-system quality and annotator access at evaluation time.

Our evaluation is conducted on a single benchmark derived from KLAR-CLC, which focuses on crosslingual knowledge consistency. The findings may not generalize to other NLP tasks such as reasoning, summarization, or open-ended generation in low-resource Indian language settings.

The prompting strategies explored in this work are evaluated on models with up to 14B parameters, which are at the smaller end of the contemporary scale. It remains an open question whether the same trends hold for larger models or models from different families. Additionally, our inference-time analysis is limited to `Llama-3.1-8B-Instruct`, and runtime characteristics may differ substantially across hardware configurations and model sizes.

All experiments use a single random seed, and we do not report confidence intervals around the reported accuracy and CLC values; quoted improvements should be interpreted as point estimates.

The code-mixed variants are generated by GPT APIs and may inherit its systematic preferences (e.g., word-class substitution patterns or English-token-ratio distribution). Despite native-speaker verification, our code-mixed distribution is therefore one of many plausible realizations of natural code-mixing rather than a sampled cross-section of it.

The interpretability and qualitative analyses in §6 are based on a single model (`Llama-3.1-8B-Instruct`) and, for the qualitative example, a single language (Hindi); mechanistic conclusions should not be inferred from these illustrations alone.

Finally, the code-mixed variants were generated for only 11 of the 18 Indian languages, constrained by available linguist coverage. Languages such as Dogri, Kannada, Konkani, Maithili, Nepali, Sindhi, and Urdu are represented only in their monolingual form, which limits the scope of code-mixed analysis for these languages.

Ethical consideration

AI-assisted writing tools were used to check grammar and improve language quality during manuscript preparation. We thank the native speakers who voluntarily contributed to the human quality verification of INDIKLAR for their careful review of the native and code-mixed variants across all 11 language settings.

Annotator Demographics. Human quality verification was conducted on a voluntary basis by 10 graduate students with native or near-native proficiency in the respective language, either as their mother tongue or through demonstrated linguistic expertise. Annotators ranged in age from 20–25 years, with a gender distribution of 7 male and 3 female participants, collectively covering 10 Indian states. Each annotator was assigned to their language of proficiency and asked to assess whether the samples appeared fluent and natural in that language. The verification task involved careful examination of the assigned language setting and took approximately 1–2 hours per language to complete.

References

Xi Ai, Mahardika Krisna Ihsani, and Min-Yen Kan. 2025. Are knowledge and reference in multilingual language models cross-lingually consistent? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4975–5011, Suzhou, China. Association for Computational Linguistics.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

David W Green and Jubin Abutalebi. 2013. Language control in bilinguals: The adaptive control hypothesis. *Journal of cognitive psychology*, 25(5):515–530.

François Grosjean. 2008. *Studying bilinguals*. Oxford University Press.

Ashim Gupta, Maitrey Mehta, Zhichao Xu, and Vivek Srikumar. 2025. Found in translation: Measuring multilingual LLM consistency as simple as translate then evaluate. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3477–3496, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2025. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. *arXiv preprint arXiv:2506.19468*.

Hongjin Kim, Jaewook Lee, Kiyoung Lee, Jong-hun Shin, Soojong Lim, and Oh-Woog Kwon. 2025. Do LLMs need inherent reasoning before reinforcement learning? a study in Korean self-correction. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 527–542, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Judith F Kroll and Erika Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of memory and language*, 33(2):149–174.

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

Yihong Liu, Mingyang Wang, François Yvon, and Hinrich Schütze. 2025. On the entity-level alignment in crosslingual consistency. *arXiv preprint arXiv:2510.10280*.

Yihong Liu, Bingyu Xiong, and Hinrich Schütze. 2026. Evaluating contextually mediated factual recall in multilingual large language models. *arXiv preprint arXiv:2601.12555*.

Nostalgebraist. 2020. interpreting gpt: the logit lens.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Dan Wang, Boxi Cao, Ning Bian, Xuanang Chen, Yaojie Lu, Hongyu Lin, Jia Zheng, Le Sun, Shanshan Jiang, Bin Dong, and Xianpei Han. 2025a. The linguistic connectivities within large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8700–8714, Vienna, Austria. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025b. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

A Experimental Settings

This section summarizes the experimental details reported in the main paper. All experiments were conducted on a single NVIDIA A100 GPU (80GB). We used HuggingFace[4] model checkpoints and employed the vLLM library (Kwon et al., 2023) for efficient batch inference and faster decoding.

Family	Parameters	Checkpoint
Llama 3.1	8B	`meta-llama/Llama-3.1-8B-Instruct`
Llama 3.2	1B	`meta-llama/Llama-3.2-1B-Instruct`
Llama 3.2	3B	`meta-llama/Llama-3.2-3B-Instruct`
Qwen 2.5	1.5B	`Qwen/Qwen2.5-1.5B-Instruct`
Qwen 2.5	7B	`Qwen/Qwen2.5-7B-Instruct`
Qwen 2.5	14B	`Qwen/Qwen2.5-14B-Instruct`
Gemma 3	1B	`google/gemma-3-1b-it`
Gemma 3	4B	`google/gemma-3-4b-it`
Gemma 3	12B	`google/gemma-3-12b-it`

Table 3: List of models used across our experiments.

Language	ISO code	Code-mixed availability
Assamese	as	✓
Bengali (Bangla)	bn	✓
Dogri	doi	
Gujarati	gu	✓
Hindi	hi	✓
Kannada	kn	
Konkani	gom	
Maithili	mai	
Malayalam	ml	✓
Marathi	mr	✓
Nepali	ne	
Odia	or	✓
Punjabi	pa	✓
Sanskrit	sa	✓
Sindhi	sd	
Tamil	ta	✓
Telugu	te	✓
Urdu	ur	

Table 4: Languages and their ISO codes. Notation: ✓indicates languages with constructed code-mixed variants, based on availability of linguists.

A.1 Evaluation Metrics

We evaluate model performance using two metrics: **Accuracy** and **Cross-Lingual Consistency (CLC)**.

Accuracy. Accuracy measures the proportion of correctly predicted facts across all evaluation instances. Since generative language models may produce partial continuations, we employ a relaxed prefix-matching (Wang et al., 2025b). A prediction is considered correct if the generated output is a non-empty case-insensitive prefix of the ground-truth answer, or vice versa. Formally, given a prediction $\hat{y}_i$ and ground-truth answer y_i, the correctness function is defined as:

$$\mathrm{Match}(\hat{y}_i, y_i) = \begin{cases} 1, & \text{if } \hat{y}_i \preceq y_i \text{ or } y_i \preceq \hat{y}_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\preceq$ denotes the prefix relation after lowercasing and whitespace normalization. The final accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\text{prefix}}(\hat{y}_i, y_i) \quad (2)$$

where N denotes the total number of evaluation samples.

Cross-Lingual Consistency (CLC). In addition to accuracy, we measure Cross-Lingual Consistency (CLC)(Qi et al., 2023) to quantify whether the model retrieves consistent knowledge across different languages.CLC is computed using the overlap ratio between the sets of correctly predicted sample indices across language pairs. Given two languages L_a and L_b, with corresponding correct prediction sets C_a and C_b, the consistency score is defined as:

$$\mathrm{CLC}(L_a, L_b) = \frac{|C_a \cap C_b|}{|C_a \cup C_b|} \quad (3)$$

The final CLC score is obtained by averaging the pairwise consistency scores across all evaluated languages. This metric captures the extent to which multilingual models preserve factual knowledge consistently across linguistic variations.

Layer-wise Rank Analysis Following Logit Lens analysis (Nostalgebraist, 2020; Wang et al., 2025b), we project intermediate hidden states at each transformer layer into the vocabulary space using the model's unembedding matrix and track the rank of the gold answer token across layers. The rank of a token is defined as the number of vocabulary tokens assigned a higher logit value by the model, where lower rank indicates stronger preference for the correct answer. Since answers may consist of multiple subword tokens, we report the minimum rank among all constituent subword tokens of the gold answer. For visualization stability, we report $\log_{10}(\mathrm{rank} + 1)$ averaged across evaluation samples.

[4] https://huggingface.co/

Variant	#Instances	Avg Query Length	Avg Answer Length
Native Indian Languages	2619	9.23	1.13
Code-Mixed Variants	2619	9.20	1.02

Table 5: Dataset statistics for INDIKLAR (open-source under CC-BY-4.0 license).

A.2 Inference Time Analysis

Table 6 reports the average per-sample end-to-end runtime across prompting strategies for all three models. Across all models, multi-step strategies (2Step-CM, 2Step-EN, 2Step-Translit) and answer-inclusive one-step variants (1Step-CM+Ans, 1Step-EN+Ans) consistently incur higher latency than the baseline. Gemma-3-12B and Qwen2.5-14B show this most clearly, with 2Step-Translit reaching 0.60s and 0.71s respectively—nearly double their baselines (0.36s and 0.34s). Both TinT variants (TinT-CM, TinT-EN) remain the most efficient across all three models: Llama-3.1-8B-Instruct runs at just 0.09s under both TinT settings, while Gemma-3-12B and Qwen2.5-14B reach at most 0.26s and 0.37s—well below their respective multi-step counterparts. These results demonstrate that TinT achieves its reasoning benefits without any additional inference overhead, making it practical for large-scale deployment.

A.3 Performance–Efficiency Tradeoffs

We further analyze the tradeoffs between prompting performance and inference efficiency by comparing the best-performing variants of 1Step, 2Step, and TinT prompting across all languages and models. For a fair comparison, we first compute the improvement of each prompting strategy over baseline for every language and model pair, and then average these improvements across languages and models. For accuracy, 1Step and 2Step prompting achieve the highest overall gains over baseline, with average improvements of $+0.1902$ and $+0.1899$ respectively, while TinT achieves a competitive improvement of $+0.1812$. For CLC evaluation, TinT achieves the strongest overall improvement ($+0.1492$), outperforming both 1Step ($+0.1416$) and 2Step ($+0.1274$). Across both evaluations, the English-guided variants emerge as the dominant settings most frequently selected as the best-performing variants. Importantly, TinT achieves these competitive gains at substantially lower inference cost. While `1Step-CM+Ans` and `2Step-CM` require approximately 0.16s and 0.17s per sample respectively, TinT variants require only 0.09s per sample. Despite this reduced runtime, TinT maintains performance comparable to stronger multi-step prompting methods, showing that TinT prompting can provide an effective balance between multilingual performance and computational efficiency.

Similar inference-time trends were also consistently observed for Gemma-3-12B-it and Qwen2.5-14B-Instruct (Appendix A.2)

A.4 Scaling Behavior and TinT Analysis

TinT's effectiveness scales clearly with model size. Smaller models such as `Llama-3.2-1B` and `Qwen2.5-1.5B` do not reliably outperform the baseline, particularly for CLC: `Llama-3.2-1B` yields positive CLC gains under `TinT-CM` for only 11 of 18 languages and under `TinT-EN` for just one, while `Qwen2.5-1.5B` improves on 15. This instability largely disappears at scale—`Llama-3.1-8B`, `Qwen2.5-7B`, and `Qwen2.5-14B` all achieve positive accuracy and CLC gains across every language under both variants. The trend is sharpest within the Qwen family, where the average `TinT-EN` accuracy gain rises from 0.062 for `Qwen2.5-1.5B` to 0.244 and 0.288 for the 7B and 14B variants.

The two variants also diverge by scale. `TinT-EN` dominates for most medium and large models, whereas `Llama-3.2-1B` favors `TinT-CM` across all 18 languages on both metrics—suggesting smaller models struggle to exploit implicit English-guided reasoning. Nonetheless, the variants remain close on average: across all models, `TinT-CM` and `TinT-EN` yield mean accuracy gains of 0.165 and 0.178, and CLC gains of 0.140 and 0.150, indicating TinT stays effective even directly in code-mixed settings.

Finally, `Gemma-3-1B` bucks the small-model trend, improving across all languages and both metrics. This suggests multilingual stability under implicit prompting depends not only on parameter count but also on model family and training characteristics.

A.5 Statistical Analysis

To assess whether the observed improvements across prompting strategies were statistically significant, we conducted paired McNemar tests over 2,619 shared evaluation samples. Both TinT-CM and TinT-EN significantly outperformed the baseline setting. TinT-CM achieved a significant improvement over the baseline ($\chi^2 = 534.75$, $p < 10^{-10}$), while TinT-EN also significantly improved

Model	Baseline	Obj-Baseline	2Step-CM	2Step-EN	2Step-Translit	1Step-CM+Ans	1Step-EN+Ans	TinT-CM	TinT-EN
Llama-3.1-8B-Instruct	0.13	0.08	0.17	0.12	0.15	0.16	0.12	0.09	0.09
Gemma-3-12B-it	0.36	0.23	0.57	0.51	0.60	0.53	0.44	0.26	0.25
Qwen-2.5-14B-Instruct	0.34	0.32	0.66	0.57	0.71	0.59	0.57	0.36	0.37

Table 6: Average per-sample end-to-end runtime (in seconds) across different prompting settings for the `meta-llama/Llama-3.1-8B-Instruct`, `google/gemma-3-12b-it`, and `Qwen/Qwen2.5-14B-Instruct` models.

over the baseline ($\chi^2 = 603.56$, $p < 10^{-10}$). Collectively, these results establish the following performance ordering:

$$\text{Baseline} < \text{TinT}.$$

A.6 Pretraining Regime

Base vs. Instruct. Table 7 compares base and instruction-tuned model variants under TinT-CM and TinT-EN settings. In the majority of cases, base and instruction-tuned variants perform comparably across both model families and all five languages. However, isolated exceptions exist: Llama-3.1-8B under TinT-CM shows a gap of up to 9.2 points on ASM (0.732 vs. 0.641), and similar but smaller divergences appear for Qwen2.5-14B on ASM under TinT-CM (0.699 vs. 0.669). These occasional gaps do not follow a consistent pattern across languages or prompting settings, suggesting they reflect task-specific variance rather than a systematic effect of instruction tuning. Overall, TinT's effectiveness relies primarily on the model's inherent multilingual representations rather than explicit instruction-following.

Model Family	Setting	AS	BN	HI	MR	TE
Llama-3.1-8B	TinT-CM	0.732	0.796	0.793	0.794	0.749
Llama-3.1-8B-Inst	TinT-CM	0.641	0.722	0.764	0.751	0.675
Llama-3.1-8B	TinT-EN	0.721	0.802	0.806	0.811	0.742
Llama-3.1-8B-Inst	TinT-EN	0.656	0.753	0.786	0.781	0.697
Qwen2.5-14B	TinT-CM	0.699	0.844	0.835	0.838	0.724
Qwen2.5-14B-Inst	TinT-CM	0.669	0.825	0.816	0.817	0.732
Qwen2.5-14B	TinT-EN	0.703	0.858	0.842	0.844	0.738
Qwen2.5-14B-Inst	TinT-EN	0.689	0.841	0.831	0.831	0.753

Table 7: Comparison between base and instruction-tuned models under Translate-in-Thought prompting settings across five representative Indian languages.

A.7 Additional Error Analysis

We observe several recurring failure modes across conversion-based prompting strategies. First, some failures occurred when the model produced little to no meaningful code-mixed conversion, in some cases simply reproducing the original native-language query or relying primarily on transliteration, thereby limiting the benefits of lexical grounding through English anchors. Although transliteration-based prompting often improved over native-script inputs, it remained less effective than proper code-mixed conversion in many cases. Second, in experiments involving intermediate transliteration, the model occasionally generated entity names that were phonetically similar to the intended entities but contained incorrect or uncommon spellings, leading to mismatches despite sounding plausible. Finally, in some cases, the model transliterated or converted only the subject/entity while leaving the remainder of the query unchanged, resulting in incomplete conversions that weakened the benefits of code-mixed prompting.

B Generalization to Contextually Mediated Knowledge Recall

To further validate our findings, we evaluate TinT on the contextually mediated knowledge recall benchmark of Liu et al. (2026), translated into our five representative Indian languages. The dataset statistics are presented in Table 8. This benchmark differs from INDIKLAR in that knowledge is accessed through naturalistic referential context rather than direct entity queries, making it a stronger test of crosslingual reasoning. Sample lengths are also considerably longer than those in KLAR-CLC. As shown in Table 9, TinT consistently improves over the baseline across all five languages and all three model families, demonstrating that the gains from Translate-in-Thought prompting generalize beyond direct knowledge recall to more contextually grounded settings.

Variant	#Instances	Avg Query Length	Avg Answer Length
Liu et al. (2026)	1742	22.00	1.02

Table 8: Dataset statistics for the contextually mediated knowledge recall benchmark.

C Language-wise Accuracy and CLC Scores on INDIKLAR

Language	English (Input)	Native (Input)	2Step-Translit (Model Intermediate) Transliteration	1Step-CM + Ans (Model Intermediate) Codemix
Assamese	What is the official language of Byelorussian Soviet Socialist Republic? ✅	বাইলৰাছিয়ান ছোভিয়েট সমাজবাদী গণৰাজ্যৰ চৰকাৰী ভাষা কি? ❌	Bairrashiyaaon Choviyee Soobheet Soomajbaadhee Gaanarajyaar Chorakari Bhaasha Ki? ❌	Bailorashiyaan Soviet Samajabadi Ganarajyaar charakari bhasha ki? ✅
Bengali	Where is Northern Territory's capital located? ✅	উত্তর টেরিটরি এর রাজধানী কোথায় অবস্থিত? ❌	uttar tiritori er rajdhani kothay abasthit? ❌	Sicily ka samrajya ki official language kya hai? ✅
Hindi	What is the official language of Kingdom of Sicily? ✅	आधिकारिक भाषा के रूप में सिसिली का साम्राज्य की कौन-सी भाषा है? ❌	ādhikārīk bhāshā ke rūp meṣ sīsīlī ka sāmrājya kī kauṇ-sī bhāshā hai? ❌	Sicily ka samrajya ki official language kya hai? ✅
Marathi	Which continent is Arab League located in? ✅	अरब लीग कोणत्या खंडात स्थित आहे? ❌	arab lig koṇatyā khaṇḍāt sthit āhe? ❌	Arab League kahaan sthit aahay? ✅

Figure 9: Qualitative comparison of different prompt representations across Assamese, Bengali, Hindi, and Marathi. Notation: The tick and cross represents correct and incorrect prediction, respectively.

		AS		BN		HI		MR		TE	
Model	**Setting**	**ACC**	**CLC**	**ACC**	**CLC**	**ACC**	**CLC**	**ACC**	**CLC**	**ACC**	**CLC**
LLaMA 3.1 8B	Baseline	0.422	0.475	0.604	0.566	0.575	0.539	0.559	0.529	0.482	0.490
	TinT	0.735	0.766	0.833	0.825	0.843	0.806	0.830	0.820	0.784	0.783
Gemma 3 12B	Baseline	0.580	0.652	0.784	0.781	0.796	0.779	0.803	0.786	0.797	0.776
	TinT	0.792	0.813	0.900	0.889	0.910	0.888	0.907	0.890	0.898	0.874
Qwen 2.5 14B	Baseline	0.520	0.575	0.706	0.686	0.695	0.671	0.729	0.682	0.565	0.589
	TinT	0.708	0.728	0.854	0.816	0.859	0.811	0.844	0.813	0.756	0.757

Table 9: Comparison between Baseline and TinT settings across Assamese (`as`), Bengali (`bn`), Hindi (`hi`), Marathi (`mr`), and Telugu (`te`). Accuracy (ACC) and Cross-Lingual Consistency (CLC) are reported for LLaMA 3.1 8B, Gemma 3 12B, and Qwen 2.5 14B Instruct models.

Lang	Strategy	Llama 3.1	Llama 3.2		Gemma 3			Qwen 2.5		
		8B	1B	3B	1B	4B	12B	1.5B	7B	14B
English	Baseline	0.842	0.744	0.805	0.672	0.763	0.870	0.833	0.879	0.840
Assamese	Baseline	0.475	0.122	0.206	0.110	0.382	0.560	0.196	0.323	0.409
	2Step-CM	+0.195	+0.106	+0.255	+0.204	+0.157	+0.131	+0.064	+0.219	+0.207
	2Step-EN	+0.223	-0.040	+0.282	+0.218	+0.169	+0.207	+0.056	+0.244	+0.287
	1Step-CM+Ans	+0.231	+0.035	+0.218	+0.176	+0.177	+0.231	+0.103	+0.227	+0.278
	1Step-EN+Ans	+0.278	-0.038	+0.223	+0.213	+0.220	+0.260	+0.082	+0.251	+0.310
	TinT-CM	+0.166	+0.123	+0.211	+0.210	+0.173	+0.204	+0.032	+0.235	+0.260
	TinT-EN	+0.181	-0.004	+0.248	+0.241	+0.183	+0.221	+0.043	+0.250	+0.281
Bengali	Baseline	0.584	0.196	0.403	0.287	0.525	0.741	0.234	0.440	0.554
	2Step-CM	+0.194	+0.056	+0.207	+0.126	+0.178	+0.118	+0.120	+0.240	+0.217
	2Step-EN	+0.239	-0.023	+0.263	+0.161	+0.245	+0.158	+0.124	+0.310	+0.302
	1Step-CM+Ans	+0.192	-0.047	+0.172	+0.086	+0.206	+0.149	+0.124	+0.261	+0.279
	1Step-EN+Ans	+0.243	-0.118	+0.134	+0.193	+0.252	+0.159	+0.087	+0.288	+0.308
	TinT-CM	+0.138	+0.023	+0.123	+0.207	+0.195	+0.144	+0.063	+0.262	+0.271
	TinT-EN	+0.169	-0.074	+0.158	+0.224	+0.200	+0.149	+0.074	+0.276	+0.288
Dogri	Baseline	0.561	0.238	0.426	0.246	0.486	0.716	0.230	0.489	0.505
	2Step-CM	+0.211	+0.157	+0.155	+0.156	+0.184	+0.096	+0.145	+0.199	+0.276
	2Step-EN	+0.254	+0.066	+0.182	+0.128	+0.176	+0.132	+0.167	+0.222	+0.330
	1Step-CM+Ans	+0.218	+0.017	+0.163	+0.077	+0.196	+0.147	+0.218	+0.222	+0.321
	1Step-EN+Ans	+0.268	-0.026	+0.196	+0.154	+0.234	+0.163	+0.134	+0.229	+0.350
	TinT-CM	+0.198	+0.175	+0.132	+0.171	+0.179	+0.143	+0.126	+0.213	+0.308
	TinT-EN	+0.221	+0.070	+0.152	+0.204	+0.207	+0.145	+0.134	+0.222	+0.326
Gujarati	Baseline	0.476	0.198	0.363	0.188	0.477	0.721	0.241	0.292	0.423
	2Step-CM	+0.254	+0.024	+0.177	+0.163	+0.207	+0.108	+0.023	+0.308	+0.341
	2Step-EN	+0.304	-0.115	+0.226	+0.176	+0.220	+0.160	+0.031	+0.370	+0.384
	1Step-CM+Ans	+0.257	-0.066	+0.106	+0.148	+0.163	+0.158	+0.055	+0.328	+0.351
	1Step-EN+Ans	+0.300	-0.121	+0.130	+0.224	+0.226	+0.167	+0.034	+0.324	+0.393
	TinT-CM	+0.220	-0.021	+0.118	+0.180	+0.178	+0.146	+0.015	+0.324	+0.339
	TinT-EN	+0.247	-0.117	+0.143	+0.217	+0.175	+0.151	+0.015	+0.331	+0.361
Hindi	Baseline	0.592	0.258	0.386	0.236	0.512	0.719	0.286	0.470	0.523
	2Step-CM	+0.174	+0.008	+0.256	+0.167	+0.207	+0.122	+0.119	+0.210	+0.273
	2Step-EN	+0.210	+0.039	+0.315	+0.229	+0.233	+0.151	+0.136	+0.278	+0.317
	1Step-CM+Ans	+0.211	-0.019	+0.248	+0.138	+0.193	+0.155	+0.079	+0.254	+0.295
	1Step-EN+Ans	+0.233	-0.081	+0.255	+0.212	+0.243	+0.154	+0.074	+0.263	+0.326
	TinT-CM	+0.172	+0.060	+0.228	+0.170	+0.201	+0.153	+0.064	+0.239	+0.292
	TinT-EN	+0.194	-0.036	+0.261	+0.205	+0.200	+0.148	+0.072	+0.257	+0.307
Kannada	Baseline	0.483	0.197	0.364	0.110	0.443	0.717	0.186	0.388	0.442
	2Step-CM	+0.276	+0.114	+0.168	+0.152	+0.194	+0.122	+0.008	+0.151	+0.250
	2Step-EN	+0.304	-0.076	+0.222	+0.169	+0.235	+0.157	+0.004	+0.219	+0.332
	1Step-CM+Ans	+0.220	-0.023	+0.140	+0.172	+0.157	+0.150	+0.117	+0.169	+0.294
	1Step-EN+Ans	+0.286	-0.085	+0.107	+0.204	+0.226	+0.169	+0.055	+0.182	+0.322
	TinT-CM	+0.187	+0.070	+0.064	+0.227	+0.171	+0.144	+0.039	+0.153	+0.259
	TinT-EN	+0.213	-0.049	+0.132	+0.271	+0.167	+0.153	+0.031	+0.170	+0.275
Konkani	Baseline	0.451	0.196	0.320	0.158	0.446	0.664	0.152	0.265	0.426
	2Step-CM	+0.288	+0.213	+0.234	+0.249	+0.201	+0.163	+0.076	+0.343	+0.331
	2Step-EN	+0.359	+0.069	+0.277	+0.263	+0.221	+0.179	+0.064	+0.395	+0.341
	1Step-CM+Ans	+0.315	-0.019	+0.187	+0.245	+0.175	+0.202	+0.068	+0.388	+0.378
	1Step-EN+Ans	+0.362	-0.033	+0.217	+0.311	+0.219	+0.211	+0.011	+0.410	+0.394
	TinT-CM	+0.285	+0.230	+0.199	+0.255	+0.183	+0.204	+0.101	+0.374	+0.362
	TinT-EN	+0.300	+0.110	+0.243	+0.287	+0.208	+0.212	+0.065	+0.393	+0.379
Maithili	Baseline	0.582	0.316	0.457	0.265	0.551	0.738	0.265	0.413	0.502
	2Step-CM	+0.197	+0.007	+0.168	+0.146	+0.171	+0.108	+0.149	+0.257	+0.279
	2Step-EN	+0.222	-0.130	+0.198	+0.140	+0.161	+0.130	+0.174	+0.305	+0.323
	1Step-CM+Ans	+0.229	-0.027	+0.159	+0.072	+0.159	+0.131	+0.118	+0.304	+0.322
	1Step-EN+Ans	+0.266	-0.084	+0.160	+0.167	+0.192	+0.140	+0.096	+0.318	+0.347
	TinT-CM	+0.191	+0.053	+0.144	+0.166	+0.164	+0.140	+0.083	+0.296	+0.316
	TinT-EN	+0.207	-0.058	+0.171	+0.211	+0.168	+0.137	+0.107	+0.309	+0.339
Malayalam	Baseline	0.528	0.153	0.354	0.049	0.480	0.721	0.270	0.400	0.516
	2Step-CM	+0.211	+0.056	+0.175	+0.100	+0.158	+0.092	-0.029	+0.164	+0.159
	2Step-EN	+0.226	-0.054	+0.197	+0.112	+0.185	+0.151	-0.021	+0.236	+0.233
	1Step-CM+Ans	+0.161	+0.029	+0.175	+0.100	+0.117	+0.132	+0.060	+0.158	+0.207
	1Step-EN+Ans	+0.240	-0.002	+0.152	+0.171	+0.163	+0.146	+0.047	+0.172	+0.252
	TinT-CM	+0.155	+0.082	+0.127	+0.144	+0.124	+0.134	+0.007	+0.148	+0.200
	TinT-EN	+0.184	+0.006	+0.165	+0.194	+0.141	+0.140	+0.034	+0.159	+0.210

Marathi	Baseline	0.592	0.242	0.399	0.303	0.537	0.764	0.310	0.443	0.579
	2Step-CM	+0.190	+0.020	+0.221	+0.095	+0.170	+0.081	+0.050	+0.224	+0.201
	2Step-EN	+0.226	+0.049	+0.259	+0.152	+0.190	+0.126	+0.059	+0.293	+0.258
	1Step-CM+Ans	+0.196	-0.005	+0.213	+0.064	+0.169	+0.124	+0.064	+0.263	+0.249
	1Step-EN+Ans	+0.234	-0.035	+0.224	+0.126	+0.221	+0.133	-0.032	+0.276	+0.279
	TinT-CM	+0.159	+0.065	+0.173	+0.118	+0.167	+0.113	+0.003	+0.252	+0.238
	TinT-EN	+0.189	-0.023	+0.211	+0.166	+0.175	+0.120	+0.001	+0.272	+0.252
Nepali	Baseline	0.556	0.227	0.352	0.156	0.518	0.714	0.289	0.439	0.364
	2Step-CM	+0.190	+0.064	+0.250	+0.268	+0.166	+0.109	+0.092	+0.239	+0.396
	2Step-EN	+0.191	-0.034	+0.197	+0.222	+0.193	+0.150	+0.133	+0.283	+0.464
	1Step-CM+Ans	+0.248	+0.014	+0.272	+0.150	+0.163	+0.153	+0.127	+0.277	+0.457
	1Step-EN+Ans	+0.265	-0.002	+0.233	+0.207	+0.219	+0.155	+0.107	+0.290	+0.480
	TinT-CM	+0.206	+0.113	+0.211	+0.290	+0.163	+0.145	+0.054	+0.271	+0.443
	TinT-EN	+0.225	-0.006	+0.263	+0.259	+0.156	+0.153	+0.081	+0.282	+0.462
Odia	Baseline	0.350	0.071	0.142	0.077	0.311	0.516	0.146	0.222	0.301
	2Step-CM	+0.248	+0.051	+0.042	+0.178	+0.170	+0.184	-0.002	+0.214	+0.324
	2Step-EN	+0.283	+0.052	+0.052	+0.167	+0.168	+0.275	-0.024	+0.250	+0.378
	1Step-CM+Ans	+0.198	+0.077	+0.085	+0.116	+0.157	+0.245	-0.002	+0.058	+0.305
	1Step-EN+Ans	+0.305	+0.018	+0.119	+0.209	+0.176	+0.283	+0.069	+0.128	+0.354
	TinT-CM	+0.188	+0.207	+0.037	+0.184	+0.178	+0.215	+0.009	+0.210	+0.291
	TinT-EN	+0.208	+0.164	+0.070	+0.172	+0.169	+0.243	+0.014	+0.251	+0.317
Punjabi	Baseline	0.476	0.187	0.344	0.118	0.417	0.654	0.206	0.339	0.494
	2Step-CM	+0.224	+0.033	+0.151	+0.051	+0.202	+0.098	+0.035	+0.172	+0.208
	2Step-EN	+0.280	-0.090	+0.218	+0.055	+0.249	+0.154	+0.036	+0.253	+0.274
	1Step-CM+Ans	+0.224	-0.049	+0.174	+0.072	+0.152	+0.168	+0.085	+0.234	+0.236
	1Step-EN+Ans	+0.291	-0.143	+0.147	+0.115	+0.181	+0.163	+0.112	+0.262	+0.297
	TinT-CM	+0.197	+0.005	+0.129	+0.081	+0.161	+0.149	+0.032	+0.222	+0.236
	TinT-EN	+0.223	-0.076	+0.144	+0.105	+0.167	+0.155	+0.050	+0.239	+0.265
Sanskrit	Baseline	0.509	0.225	0.312	0.190	0.470	0.695	0.215	0.377	0.517
	2Step-CM	+0.236	+0.120	+0.229	+0.202	+0.165	+0.125	+0.128	+0.272	+0.255
	2Step-EN	+0.282	+0.005	+0.264	+0.189	+0.151	+0.158	+0.157	+0.320	+0.296
	1Step-CM+Ans	+0.256	-0.004	+0.177	+0.129	+0.171	+0.159	+0.112	+0.328	+0.312
	1Step-EN+Ans	+0.311	-0.023	+0.281	+0.191	+0.240	+0.176	+0.065	+0.340	+0.334
	TinT-CM	+0.226	+0.147	+0.163	+0.274	+0.171	+0.154	+0.076	+0.307	+0.288
	TinT-EN	+0.261	+0.033	+0.195	+0.271	+0.190	+0.158	+0.076	+0.323	+0.313
Sindhi	Baseline	0.478	0.172	0.249	0.129	0.416	0.540	0.188	0.369	0.509
	2Step-CM	+0.121	+0.069	+0.079	+0.181	+0.119	+0.110	+0.101	+0.121	+0.074
	2Step-EN	+0.192	-0.055	+0.105	+0.152	+0.145	+0.208	+0.045	+0.081	+0.073
	1Step-CM+Ans	+0.193	-0.007	+0.156	+0.142	+0.123	+0.207	+0.143	+0.183	+0.128
	1Step-EN+Ans	+0.247	-0.066	+0.150	+0.233	+0.175	+0.249	+0.164	+0.198	+0.153
	TinT-CM	+0.153	+0.112	+0.117	+0.205	+0.141	+0.188	+0.127	+0.178	+0.114
	TinT-EN	+0.187	+0.008	+0.140	+0.252	+0.153	+0.204	+0.105	+0.189	+0.148
Tamil	Baseline	0.440	0.181	0.329	0.285	0.488	0.700	0.200	0.303	0.392
	2Step-CM	+0.231	-0.001	+0.166	+0.139	+0.190	+0.067	-0.018	+0.165	+0.219
	2Step-EN	+0.271	-0.049	+0.219	+0.149	+0.226	+0.165	-0.024	+0.194	+0.277
	1Step-CM+Ans	+0.231	+0.004	+0.199	+0.089	+0.185	+0.164	+0.054	+0.172	+0.276
	1Step-EN+Ans	+0.289	-0.080	+0.148	+0.186	+0.250	+0.175	+0.089	+0.207	+0.297
	TinT-CM	+0.194	+0.027	+0.084	+0.146	+0.187	+0.141	+0.010	+0.179	+0.255
	TinT-EN	+0.218	-0.082	+0.134	+0.194	+0.181	+0.142	+0.010	+0.198	+0.266
Telugu	Baseline	0.509	0.181	0.385	0.192	0.499	0.745	0.189	0.473	0.515
	2Step-CM	+0.216	+0.100	+0.189	+0.190	+0.209	+0.108	+0.011	+0.088	+0.214
	2Step-EN	+0.279	-0.012	+0.225	+0.179	+0.233	+0.151	+0.011	+0.139	+0.279
	1Step-CM+Ans	+0.182	+0.019	+0.151	+0.106	+0.195	+0.144	+0.093	+0.089	+0.242
	1Step-EN+Ans	+0.261	+0.018	+0.134	+0.234	+0.228	+0.153	+0.126	+0.107	+0.276
	TinT-CM	+0.166	+0.084	+0.107	+0.198	+0.168	+0.128	+0.020	+0.074	+0.217
	TinT-EN	+0.188	-0.013	+0.158	+0.286	+0.192	+0.132	+0.070	+0.096	+0.238
Urdu	Baseline	0.570	0.265	0.407	0.137	0.483	0.688	0.251	0.485	0.590
	2Step-CM	+0.139	-0.058	+0.138	+0.124	+0.152	+0.081	+0.116	+0.096	+0.092
	2Step-EN	+0.204	-0.092	+0.239	+0.134	+0.195	+0.155	+0.107	+0.172	+0.180
	1Step-CM+Ans	+0.190	-0.078	+0.163	+0.093	+0.132	+0.138	+0.126	+0.158	+0.174
	1Step-EN+Ans	+0.233	-0.140	+0.151	+0.172	+0.203	+0.147	+0.140	+0.189	+0.189
	TinT-CM	+0.137	-0.019	+0.112	+0.168	+0.125	+0.132	+0.120	+0.153	+0.149
	TinT-EN	+0.181	-0.100	+0.166	+0.205	+0.145	+0.133	+0.128	+0.170	+0.160

Table 10: Consolidated multilingual evaluation accuracy results across all model families and prompting strategies. Scores are reported as accuracy in the range $[0, 1]$.

Lang	Strategy	Llama 3.1	Llama 3.2		Gemma 3			Qwen 2.5		
		8B	1B	3B	1B	4B	12B	1.5B	7B	14B
Assamese	Baseline	0.518	0.263	0.322	0.197	0.463	0.637	0.362	0.408	0.480
	2Step-CM	+0.170	-0.000	+0.190	+0.095	+0.113	+0.063	-0.024	+0.143	+0.142
	2Step-EN	+0.184	-0.133	+0.189	+0.094	+0.118	+0.135	-0.025	+0.167	+0.209
	1Step-CM+Ans	+0.202	-0.003	+0.158	+0.112	+0.156	+0.174	-0.057	+0.172	+0.218
	1Step-EN+Ans	+0.254	-0.113	+0.163	+0.140	+0.167	+0.201	-0.074	+0.195	+0.248
	TinT-CM	+0.191	+0.032	+0.186	+0.135	+0.140	+0.153	-0.020	+0.192	+0.212
	TinT-EN	+0.204	-0.070	+0.210	+0.153	+0.164	+0.165	-0.012	+0.205	+0.230
Bengali	Baseline	0.590	0.338	0.456	0.277	0.535	0.740	0.351	0.461	0.548
	2Step-CM	+0.152	-0.042	+0.129	+0.084	+0.140	+0.064	+0.040	+0.149	+0.168
	2Step-EN	+0.194	-0.142	+0.151	+0.078	+0.166	+0.124	+0.029	+0.207	+0.238
	1Step-CM+Ans	+0.169	-0.079	+0.106	+0.085	+0.162	+0.133	-0.006	+0.196	+0.247
	1Step-EN+Ans	+0.224	-0.190	+0.097	+0.130	+0.190	+0.150	-0.043	+0.215	+0.268
	TinT-CM	+0.166	-0.041	+0.127	+0.131	+0.158	+0.123	+0.039	+0.202	+0.236
	TinT-EN	+0.186	-0.120	+0.149	+0.160	+0.174	+0.132	+0.052	+0.215	+0.247
Dogri	Baseline	0.572	0.323	0.454	0.268	0.516	0.747	0.357	0.449	0.534
	2Step-CM	+0.182	+0.014	+0.140	+0.116	+0.159	+0.046	+0.049	+0.175	+0.202
	2Step-EN	+0.217	-0.106	+0.136	+0.078	+0.157	+0.092	+0.038	+0.228	+0.252
	1Step-CM+Ans	+0.188	-0.025	+0.121	+0.113	+0.174	+0.118	+0.028	+0.214	+0.264
	1Step-EN+Ans	+0.247	-0.106	+0.098	+0.142	+0.193	+0.136	-0.032	+0.239	+0.288
	TinT-CM	+0.196	+0.051	+0.137	+0.151	+0.167	+0.108	+0.069	+0.228	+0.260
	TinT-EN	+0.219	-0.037	+0.147	+0.172	+0.187	+0.111	+0.086	+0.244	+0.271
Gujarati	Baseline	0.537	0.290	0.442	0.256	0.507	0.746	0.331	0.402	0.496
	2Step-CM	+0.199	-0.005	+0.124	+0.091	+0.166	+0.057	+0.004	+0.199	+0.232
	2Step-EN	+0.230	-0.151	+0.140	+0.093	+0.178	+0.116	-0.004	+0.248	+0.274
	1Step-CM+Ans	+0.199	-0.044	+0.078	+0.102	+0.169	+0.124	-0.020	+0.230	+0.270
	1Step-EN+Ans	+0.257	-0.148	+0.077	+0.149	+0.199	+0.143	-0.068	+0.239	+0.305
	TinT-CM	+0.209	-0.031	+0.124	+0.126	+0.174	+0.117	+0.010	+0.246	+0.266
	TinT-EN	+0.226	-0.115	+0.135	+0.166	+0.184	+0.125	+0.024	+0.261	+0.283
Hindi	Baseline	0.569	0.315	0.422	0.263	0.530	0.743	0.323	0.438	0.515
	2Step-CM	+0.176	+0.001	+0.164	+0.119	+0.158	+0.060	+0.088	+0.175	+0.217
	2Step-EN	+0.210	-0.101	+0.190	+0.121	+0.168	+0.113	+0.089	+0.239	+0.273
	1Step-CM+Ans	+0.194	-0.027	+0.152	+0.116	+0.166	+0.123	+0.044	+0.229	+0.272
	1Step-EN+Ans	+0.239	-0.103	+0.133	+0.157	+0.194	+0.137	-0.003	+0.253	+0.302
	TinT-CM	+0.190	+0.050	+0.168	+0.159	+0.170	+0.115	+0.108	+0.236	+0.270
	TinT-EN	+0.210	-0.038	+0.180	+0.174	+0.176	+0.121	+0.126	+0.255	+0.284
Kannada	Baseline	0.565	0.314	0.464	0.191	0.527	0.735	0.357	0.445	0.515
	2Step-CM	+0.177	-0.002	+0.102	+0.082	+0.127	+0.067	-0.065	+0.109	+0.162
	2Step-EN	+0.206	-0.147	+0.123	+0.088	+0.148	+0.122	-0.078	+0.166	+0.227
	1Step-CM+Ans	+0.153	-0.033	+0.076	+0.123	+0.137	+0.131	-0.044	+0.151	+0.227
	1Step-EN+Ans	+0.228	-0.161	+0.050	+0.150	+0.167	+0.151	-0.088	+0.172	+0.250
	TinT-CM	+0.170	+0.012	+0.074	+0.150	+0.144	+0.125	-0.024	+0.161	+0.207
	TinT-EN	+0.182	-0.096	+0.109	+0.189	+0.146	+0.132	-0.034	+0.174	+0.218
Konkani	Baseline	0.511	0.298	0.414	0.234	0.487	0.700	0.331	0.380	0.467
	2Step-CM	+0.208	+0.025	+0.159	+0.117	+0.168	+0.097	-0.038	+0.225	+0.256
	2Step-EN	+0.257	-0.113	+0.167	+0.115	+0.172	+0.136	-0.060	+0.252	+0.274
	1Step-CM+Ans	+0.236	-0.039	+0.128	+0.117	+0.179	+0.157	-0.083	+0.265	+0.311
	1Step-EN+Ans	+0.289	-0.106	+0.119	+0.155	+0.200	+0.176	-0.134	+0.285	+0.326
	TinT-CM	+0.244	+0.057	+0.161	+0.147	+0.175	+0.153	+0.031	+0.270	+0.305
	TinT-EN	+0.259	-0.039	+0.162	+0.180	+0.193	+0.162	+0.012	+0.293	+0.320
Maithili	Baseline	0.603	0.353	0.485	0.286	0.547	0.753	0.382	0.458	0.529
	2Step-CM	+0.158	+0.006	+0.115	+0.114	+0.153	+0.064	+0.048	+0.166	+0.214
	2Step-EN	+0.185	-0.161	+0.120	+0.086	+0.151	+0.102	+0.028	+0.225	+0.263
	1Step-CM+Ans	+0.166	-0.047	+0.107	+0.111	+0.159	+0.120	-0.004	+0.223	+0.271
	1Step-EN+Ans	+0.218	-0.117	+0.087	+0.145	+0.187	+0.137	-0.047	+0.245	+0.293
	TinT-CM	+0.174	+0.041	+0.121	+0.158	+0.162	+0.116	+0.073	+0.232	+0.266
	TinT-EN	+0.190	-0.050	+0.110	+0.185	+0.177	+0.120	+0.084	+0.247	+0.280
Malayalam	Baseline	0.555	0.294	0.430	0.101	0.529	0.728	0.319	0.397	0.517
	2Step-CM	+0.176	-0.025	+0.130	+0.089	+0.109	+0.059	+0.021	+0.177	+0.145
	2Step-EN	+0.193	-0.166	+0.133	+0.100	+0.128	+0.121	+0.008	+0.229	+0.213
	1Step-CM+Ans	+0.160	-0.024	+0.112	+0.117	+0.117	+0.121	+0.005	+0.197	+0.206
	1Step-EN+Ans	+0.237	-0.114	+0.094	+0.175	+0.145	+0.144	-0.011	+0.220	+0.241
	TinT-CM	+0.183	+0.005	+0.127	+0.154	+0.123	+0.124	+0.044	+0.208	+0.208
	TinT-EN	+0.204	-0.069	+0.149	+0.193	+0.137	+0.126	+0.054	+0.215	+0.213
Marathi	Baseline	0.577	0.301	0.451	0.288	0.547	0.760	0.338	0.459	0.552
	2Step-CM	+0.174	+0.014	+0.130	+0.091	+0.130	+0.053	+0.062	+0.165	+0.179

	2Step-EN	+0.206	-0.079	+0.152	+0.085	+0.147	+0.106	+0.057	+0.220	+0.233
	1Step-CM+Ans	+0.189	-0.007	+0.128	+0.105	+0.156	+0.118	+0.040	+0.212	+0.241
	1Step-EN+Ans	+0.238	-0.083	+0.102	+0.142	+0.182	+0.135	-0.031	+0.231	+0.266
	TinT-CM	+0.189	+0.046	+0.144	+0.140	+0.152	+0.107	+0.094	+0.221	+0.235
	TinT-EN	+0.205	-0.024	+0.157	+0.165	+0.167	+0.114	+0.105	+0.238	+0.246
Nepali	Baseline	0.555	0.326	0.431	0.248	0.521	0.737	0.357	0.449	0.442
	2Step-CM	+0.174	+0.012	+0.138	+0.135	+0.151	+0.060	+0.060	+0.171	+0.282
	2Step-EN	+0.186	-0.120	+0.113	+0.098	+0.174	+0.114	+0.054	+0.220	+0.337
	1Step-CM+Ans	+0.204	-0.032	+0.138	+0.119	+0.158	+0.134	+0.027	+0.210	+0.342
	1Step-EN+Ans	+0.251	-0.111	+0.108	+0.149	+0.195	+0.145	+0.004	+0.225	+0.365
	TinT-CM	+0.194	+0.043	+0.144	+0.175	+0.159	+0.123	+0.074	+0.217	+0.333
	TinT-EN	+0.213	-0.033	+0.166	+0.184	+0.171	+0.133	+0.079	+0.234	+0.351
Odia	Baseline	0.440	0.189	0.241	0.138	0.357	0.587	0.299	0.352	0.397
	2Step-CM	+0.178	-0.029	-0.009	+0.072	+0.174	+0.122	-0.095	+0.126	+0.239
	2Step-EN	+0.203	-0.063	-0.014	+0.065	+0.165	+0.209	-0.110	+0.162	+0.280
	1Step-CM+Ans	+0.131	+0.012	-0.019	+0.068	+0.176	+0.204	-0.104	-0.021	+0.247
	1Step-EN+Ans	+0.230	-0.058	+0.035	+0.120	+0.177	+0.235	-0.075	+0.053	+0.291
	TinT-CM	+0.147	+0.056	-0.017	+0.102	+0.184	+0.177	-0.068	+0.153	+0.241
	TinT-EN	+0.176	+0.004	+0.015	+0.086	+0.181	+0.198	-0.061	+0.183	+0.258
Punjabi	Baseline	0.528	0.281	0.433	0.183	0.485	0.708	0.289	0.410	0.509
	2Step-CM	+0.173	-0.054	+0.080	-0.008	+0.136	+0.031	-0.003	+0.113	+0.172
	2Step-EN	+0.220	-0.146	+0.117	-0.002	+0.163	+0.095	-0.012	+0.184	+0.235
	1Step-CM+Ans	+0.179	-0.028	+0.103	+0.058	+0.143	+0.116	+0.028	+0.178	+0.226
	1Step-EN+Ans	+0.251	-0.192	+0.068	+0.049	+0.150	+0.123	-0.004	+0.206	+0.272
	TinT-CM	+0.184	-0.086	+0.121	+0.041	+0.132	+0.104	+0.017	+0.185	+0.231
	TinT-EN	+0.206	-0.099	+0.135	+0.097	+0.166	+0.112	+0.075	+0.201	+0.249
Sanskrit	Baseline	0.536	0.314	0.386	0.254	0.524	0.733	0.282	0.436	0.530
	2Step-CM	+0.198	+0.010	+0.172	+0.109	+0.123	+0.064	+0.081	+0.183	+0.191
	2Step-EN	+0.236	-0.105	+0.178	+0.090	+0.132	+0.111	+0.083	+0.226	+0.243
	1Step-CM+Ans	+0.208	-0.023	+0.141	+0.124	+0.153	+0.125	+0.040	+0.232	+0.261
	1Step-EN+Ans	+0.272	-0.111	+0.162	+0.144	+0.186	+0.148	+0.007	+0.251	+0.285
	TinT-CM	+0.218	+0.044	+0.170	+0.171	+0.151	+0.120	+0.106	+0.241	+0.249
	TinT-EN	+0.239	-0.034	+0.186	+0.182	+0.168	+0.129	+0.127	+0.257	+0.265
Sindhi	Baseline	0.493	0.268	0.337	0.211	0.452	0.606	0.263	0.394	0.474
	2Step-CM	+0.121	-0.034	+0.048	+0.081	+0.108	+0.063	+0.028	+0.105	+0.118
	2Step-EN	+0.176	-0.155	+0.046	+0.048	+0.121	+0.148	-0.001	+0.082	+0.115
	1Step-CM+Ans	+0.169	-0.056	+0.086	+0.063	+0.128	+0.168	+0.035	+0.162	+0.186
	1Step-EN+Ans	+0.235	-0.141	+0.070	+0.128	+0.158	+0.204	+0.025	+0.173	+0.211
	TinT-CM	+0.162	-0.004	+0.086	+0.107	+0.129	+0.149	+0.053	+0.167	+0.180
	TinT-EN	+0.192	-0.087	+0.087	+0.141	+0.145	+0.163	+0.033	+0.183	+0.202
Tamil	Baseline	0.529	0.284	0.428	0.272	0.518	0.731	0.280	0.375	0.442
	2Step-CM	+0.160	-0.034	+0.110	+0.092	+0.138	+0.020	+0.001	+0.124	+0.172
	2Step-EN	+0.180	-0.111	+0.123	+0.086	+0.165	+0.118	-0.006	+0.150	+0.216
	1Step-CM+Ans	+0.163	-0.031	+0.099	+0.098	+0.156	+0.127	-0.005	+0.141	+0.238
	1Step-EN+Ans	+0.229	-0.155	+0.068	+0.139	+0.184	+0.144	+0.004	+0.176	+0.258
	TinT-CM	+0.176	-0.033	+0.082	+0.133	+0.155	+0.116	+0.037	+0.162	+0.227
	TinT-EN	+0.188	-0.126	+0.081	+0.164	+0.166	+0.119	+0.036	+0.180	+0.239
Telugu	Baseline	0.555	0.324	0.458	0.255	0.533	0.747	0.298	0.408	0.530
	2Step-CM	+0.163	-0.008	+0.118	+0.104	+0.141	+0.052	+0.003	+0.168	+0.171
	2Step-EN	+0.212	-0.135	+0.119	+0.078	+0.163	+0.117	-0.005	+0.212	+0.226
	1Step-CM+Ans	+0.148	-0.036	+0.097	+0.096	+0.161	+0.125	-0.006	+0.199	+0.217
	1Step-EN+Ans	+0.238	-0.145	+0.074	+0.155	+0.185	+0.144	+0.003	+0.220	+0.246
	TinT-CM	+0.186	-0.003	+0.114	+0.133	+0.155	+0.114	+0.026	+0.206	+0.212
	TinT-EN	+0.199	-0.071	+0.133	+0.173	+0.179	+0.124	+0.053	+0.222	+0.228
Urdu	Baseline	0.563	0.315	0.438	0.230	0.523	0.715	0.351	0.455	0.531
	2Step-CM	+0.132	-0.078	+0.113	+0.054	+0.122	+0.043	+0.014	+0.105	+0.128
	2Step-EN	+0.190	-0.158	+0.135	+0.054	+0.139	+0.120	+0.010	+0.163	+0.206
	1Step-CM+Ans	+0.168	-0.084	+0.094	+0.062	+0.124	+0.119	-0.008	+0.165	+0.218
	1Step-EN+Ans	+0.225	-0.152	+0.090	+0.116	+0.153	+0.136	-0.034	+0.191	+0.241
	TinT-CM	+0.154	-0.026	+0.107	+0.100	+0.120	+0.107	+0.008	+0.171	+0.197
	TinT-EN	+0.184	-0.113	+0.112	+0.130	+0.141	+0.109	+0.020	+0.191	+0.217

Table 11: Consolidated Cross-Lingual Consistency (CLC) results across all model families and prompting strategies. Scores denote the average pairwise Jaccard overlap between the sets of correctly predicted factual indices across multilingual language pairs.

D Prompting Strategies

This section presents the prompts used in our experiments. Variables enclosed in angle-brackets (SOURCE_LANG, SOURCE_SCRIPT, TARGET_LANG) are runtime configuration parameters passed via command-line arguments. Placeholders written in SMALL-CAPS (e.g., QUESTION, ANSWER) represent concrete values filled in at runtime from the dataset.

D.1 Baseline Prompting

We evaluate models under a 3-shot cloze-completion setting. Each prompt is constructed from a relation-specific template containing two placeholders: <subject>, filled with the entity name, and <mask>, which is removed to leave an open-ended sentence for the model to complete. For each test instance, three demonstrations are sampled from the same relation and language, and prepended as few-shot pairs. The test item is appended without its target, and the model generates the answer as a greedy continuation. Each demonstration line is formed by filling <subject> with the entity name, stripping <mask>, and appending the gold object with a single leading space. The test item omits the target entirely, prompting the model to continue the incomplete sentence. A prediction is accepted as correct if it is a case-insensitive prefix of the gold answer, or vice versa.

A second variant extends this setup by appending the set of candidate answers directly to the test item, followed by an Answer: marker, while keeping the three few-shot demonstrations in the same plain completion format. When candidate answers are available, the test item is constructed by appending the candidate list and the Answer: marker to the filled template. If no candidates are provided, the prompt falls back to the plain completion format described above.

Baseline Prompts

```
[template with <subject_1> filled] <
    answer_1>
[template with <subject_2> filled] <
    answer_2>
[template with <subject_3> filled] <
    answer_3>
[template with <subject_4> filled]
```

[Baseline]

```
[template with <subject_1> filled] <
    answer_1>
[template with <subject_2> filled] <
    answer_2>
[template with <subject_3> filled] <
    answer_3>
[template with <subject_4> filled]
Candidates: <candidate_1>, <candidate_2>,
    ..., <candidate_n>
Answer:
```

[Baseline + Candidates]

D.2 Two-Step Prompting Strategy

The 2 Step strategy decomposes each query into two sequential inference passes. Stage 1 reformulates the source-language question into an intermediate form; Stage 2 receives that reformulated question together with three in-context demonstrations and selects the correct answer from the candidate list. Both stages return structured JSON output. We evaluate three reformulation strategies under this framework.

D.2.1 English Reformulation

Stage 1 translates the source-language question into English. Stage 2 receives the English question and answers in the source script.

Stage 1 — Translation.

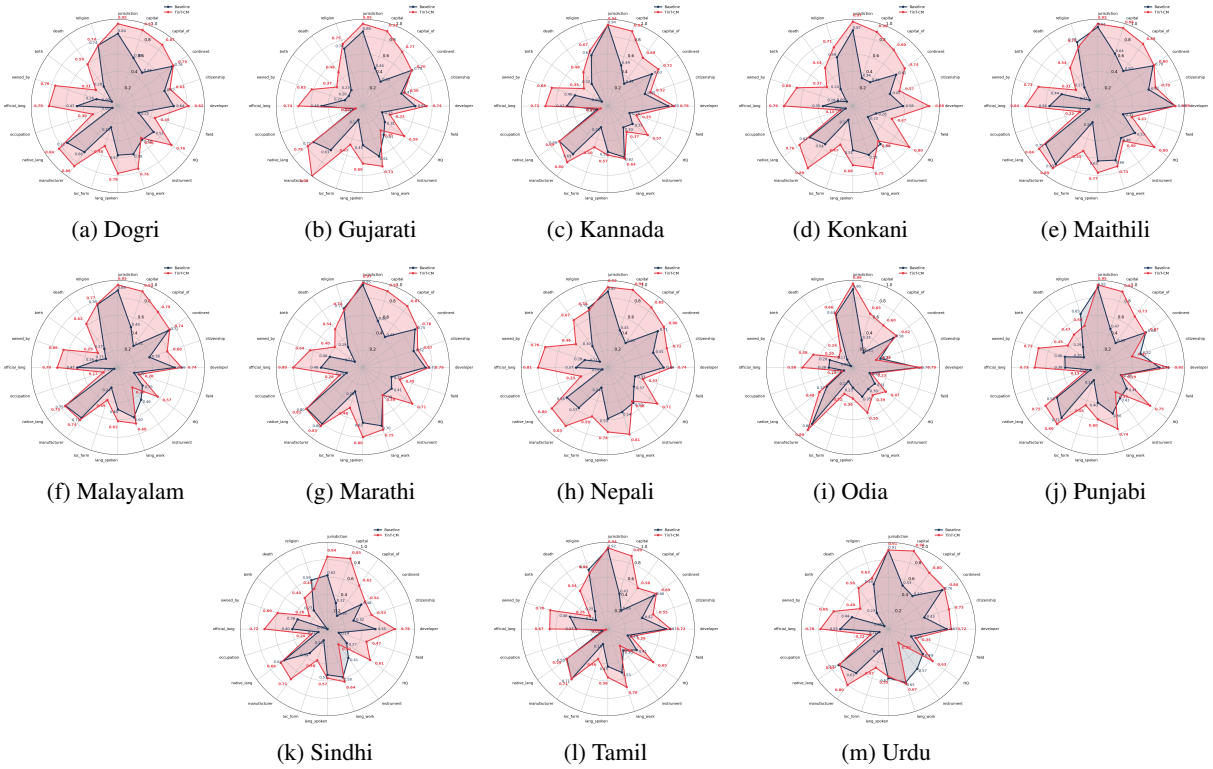
English Reformulation — Stage 1

```
You are a translator.
Translate the following question from <
    SOURCE_LANG> (<SOURCE_SCRIPT>)
into English.

Output ONLY: {"translation": "..."}

Q: <QUESTION>
Output:
```

Stage 2 — Answer. The English question from Stage 1 is presented alongside three source-language demonstrations, each showing the original question, its candidates, and the gold answer.

English Reformulation — Stage 2

```
You are answering a factual question in
    English.
Select the correct answer from the
    candidates and write
it in <SOURCE_LANG> (<SOURCE_SCRIPT>).

Output ONLY: {"answer": "..."}
```



(a) Dogri (b) Gujarati (c) Kannada (d) Konkani (e) Maithili

(f) Malayalam (g) Marathi (h) Nepali (i) Odia (j) Punjabi

(k) Sindhi (l) Tamil (m) Urdu

Figure 10: Relation-wise performance comparison of TinT prompting and baseline prompting across the remaining 13 languages using `meta-llama/Llama-3.1-8B-Instruct`.

```
Rules:
- Answer MUST be from the candidate list
- Answer MUST be in <SOURCE_SCRIPT> script

Q: <QUESTION_1>
Candidates: <CANDIDATES_1>
Output: {"answer": "<ANSWER_1>"}

Q: <QUESTION_2>
Candidates: <CANDIDATES_2>
Output: {"answer": "<ANSWER_2>"}

Q: <QUESTION_3>
Candidates: <CANDIDATES_3>
Output: {"answer": "<ANSWER_3>"}

Q: <TRANSLATED_QUESTION>
Candidates: <CANDIDATES>
Output:
```

D.2.2 Code-Mixed Reformulation

Stage 1 converts the source-language question into a code-mixed form, preserving the source-language grammatical structure while replacing content words with English equivalents in Roman script. Stage 2 answers the code-mixed question in the source script.

Stage 1 — Code-Mixed Conversion.

Code-Mixed Reformulation — Stage 1

```
You are a STRICT translator performing
    code-mixing.

Convert this <SOURCE_LANG> (<SOURCE_SCRIPT
    >) question into
<TARGET_LANG> (Roman script), preserving
    the original structure.

Definition of code-mixing:
- Keep <SOURCE_LANG> function words in
    romanized form
- Replace ONLY content words with English
    equivalents
- Preserve word order exactly
- Output MUST NOT be a fully fluent
    English sentence

Example (Hinglish):
  Input:     Bharat ki rajdhani kya hai?
  Output:    Bharat ki capital kya hai?

Output ONLY: {"translation": "<code-mixed
    question>"}

Q: <QUESTION>
Output:
```

Stage 2 — Answer. Stage 2 is identical in structure to the English reformulation Stage 2, with

three source-language demonstrations and the code-mixed question as the test item.

Code-Mixed Reformulation — Stage 2

```
You are answering a factual question in <
    TARGET_LANG>.
This uses <SOURCE_LANG> grammar with
    English vocabulary
(Roman script).
Output ONLY: {"answer": "..."} in <
    SOURCE_SCRIPT> script.

Rules:
- Answer MUST be from the candidate list
- Answer MUST be in <SOURCE_SCRIPT> script

Q: <QUESTION_1>
Candidates: <CANDIDATES_1>
Output: {"answer": "<ANSWER_1>"}

Q: <QUESTION_2>
Candidates: <CANDIDATES_2>
Output: {"answer": "<ANSWER_2>"}

Q: <QUESTION_3>
Candidates: <CANDIDATES_3>
Output: {"answer": "<ANSWER_3>"}

Q: <CODE_MIXED_QUESTION>
Candidates: <CANDIDATES>
Output:
```

D.2.3 Transliteration-Based Reformulation

Stage 1 phonetically converts the <source-language> question from its native script into Roman script without translating any words. Stage 2 answers the romanized question in the source script. This variant isolates the effect of script conversion alone, keeping all vocabulary in the source language.

Stage 1 — Transliteration.

Transliteration — Stage 1

```
You are a transliterator. Convert the
    following <SOURCE_LANG>
question from <SOURCE_SCRIPT> into Roman/
    Latin script.

- Rewrite every word phonetically in Roman
     script
- Do NOT translate or replace any word
- Preserve all words and their order
    exactly

Output ONLY: {"translation": "<romanized
    question>"}

Q: <QUESTION>
Output:
```

Stage 2 — Answer. Stage 2 is identical in structure to the code-mixed Stage 2, using three source-language demonstrations and the romanized question as the test item.

One-Step Prompting Strategy The one-step strategy performs reformulation and answering in a *single* inference pass. We evaluate four variants under this framework, organized by reformulation type.

D.2.4 English Setting

TinT-EN. The model is instructed to internally translate the source-language question into English before answering. The translation never appears in the output; only the answer is returned.

TinT-EN

```
You are answering factual questions
    written in <SOURCE_LANG>
(<SOURCE_SCRIPT> script).

You MUST internally translate the question
     into <TARGET_LANG>
to fully understand its meaning. This
    translation is ONLY
for internal reasoning --- do NOT output
    it.

Select the correct answer from the
    candidates.
Answer MUST be from the candidate list.
Answer MUST be in <SOURCE_SCRIPT> script
    only.

Output ONLY: {"answer": "<candidate>"}

Q: <QUESTION_1>
Candidates: <CANDIDATES_1>
Output: {"answer": "<ANSWER_1>"}

Q: <QUESTION_2>
Candidates: <CANDIDATES_2>
Output: {"answer": "<ANSWER_2>"}

Q: <QUESTION_3>
Candidates: <CANDIDATES_3>
Output: {"answer": "<ANSWER_3>"}

Q: <QUESTION>
Candidates: <CANDIDATES>
Output:
```

1Step-EN+Ans. The model outputs both an English translation and the <source-script> answer

in a single JSON object. The translation field is logged for analysis; only the answer field is evaluated.

1Step-EN+Ans

```
You are answering factual questions
    written in <SOURCE_LANG>
(<SOURCE_SCRIPT> script).

Translate the question into <TARGET_LANG>,
     then select the
correct answer from the candidates.

Output ONLY:
{"translation": "<English sentence>", "
    answer": "<candidate>"}

Rules:
- Translation MUST be a fluent <
    TARGET_LANG> sentence
- Translation MUST NOT contain <
    SOURCE_SCRIPT> text
- Answer MUST be chosen from the candidate
     list
- Answer MUST be in <SOURCE_SCRIPT> script
     only

Q: <QUESTION_1>
Candidates: <CANDIDATES_1>
Output: {"answer": "<ANSWER_1>"}

Q: <QUESTION_2>
Candidates: <CANDIDATES_2>
Output: {"answer": "<ANSWER_2>"}

Q: <QUESTION_3>
Candidates: <CANDIDATES_3>
Output: {"answer": "<ANSWER_3>"}

Q: <QUESTION>
Candidates: <CANDIDATES>
Output:
```

D.2.5 Code-Mixed Setting

TinT-CM The model is instructed to internally perform a code-mixed transformation before answering. This intermediate step is never output; only the answer appears in the response.

TinT-CM

```
You are answering factual questions
    written in SOURCE_LANG
(SOURCE_SCRIPT script).

Before answering, internally convert the
    question to
TARGET_LANG by keeping SOURCE_LANG grammar
     and replacing
only content words with English (Roman
    script). Do NOT
output this step.

Example of the internal transformation:
  Input:    Bharat ki rajdhani kya hai?
  Internal: Bharat ki capital kya hai?

Select the correct answer from the
    candidates.
Answer MUST be from the candidate list.
Answer MUST be in SOURCE_SCRIPT script
    only.

Output ONLY: {"answer": "<candidate>"}

Q: <QUESTION_1>
Candidates: <CANDIDATES_1>
Output: {"answer": "<ANSWER_1>"}

Q: <QUESTION_2>
Candidates: <CANDIDATES_2>
Output: {"answer": "<ANSWER_2>"}

Q: <QUESTION_3>
Candidates: <CANDIDATES_3>
Output: {"answer": "<ANSWER_3>"}

Q: <QUESTION>
Candidates: <CANDIDATES>
Output:
```

1Step-CM+Ans. The model outputs both a code-mixed transformation and the source-script answer in a single JSON object. The code-mixed field is logged for analysis; evaluation uses only the answer field.

1Step-CM+Ans

```
You are answering factual questions
    written in <SOURCE_LANG>
(<SOURCE_SCRIPT> script).

Convert the question to code-mixed <
    TARGET_LANG> (<SOURCE_LANG>
grammar + English content words in Roman
    script), then
select the correct answer from the
    candidates.

Example:
  Input:      Bharat ki rajdhani kya hai?
  Code-mixed: Bharat ki capital kya hai?

Output ONLY:
{"translation": "<code-mixed question>", "
    answer": "<candidate>"}

Rules:
```

```
- Translation MUST preserve <SOURCE_LANG>
    word order
- Translation MUST NOT be fully fluent
    English
- Answer MUST be chosen from the candidate
     list
- Answer MUST be in <SOURCE_SCRIPT> script
     only

Q: <QUESTION_1>
Candidates: <CANDIDATES_1>
Output: {"answer": "<ANSWER_1>"}

Q: <QUESTION_2>
Candidates: <CANDIDATES_2>
Output: {"answer": "<ANSWER_2>"}

Q: <QUESTION_3>
Candidates: <CANDIDATES_3>
Output: {"answer": "<ANSWER_3>"}

Q: <QUESTION>
Candidates: <CANDIDATES>
Output:
```

CoT Prompts

```
Answer the following question.
Output JSON with 'reasoning' (1-2
    sentences) and 'answer'.

Q: [template with <subject_1> filled]
Output: {"reasoning": "...", "answer": "<
    answer_1>"}

Q: [template with <subject_2> filled]
Output: {"reasoning": "...", "answer": "<
    answer_2>"}

Q: [template with <subject_3> filled]
Output: {"reasoning": "...", "answer": "<
    answer_3>"}

Q: [template with <subject_4> filled]
Output:
```

[Baseline + CoT]

```
Answer the following question.
Output JSON with 'reasoning' (1-2
    sentences) and 'answer'.

Q: [template with <subject_1> filled]
Output: {"reasoning": "...", "answer": "<
    answer_1>"}

Q: [template with <subject_2> filled]
Output: {"reasoning": "...", "answer": "<
    answer_2>"}

Q: [template with <subject_3> filled]
Output: {"reasoning": "...", "answer": "<
    answer_3>"}

Q: [template with <subject_4> filled]
Candidates: <candidate_1>, <candidate_2>,
    ..., <candidate_n>
Output:
```

[Baseline + Candidates + CoT]