IndicRipple: Measuring Cross-Lingual Ripple Effects of Knowledge Editing in Indic Languages

Anonymous ACL submission

Abstract

Large language models can be updated after training by modifying parameters, conditioning on edit demonstrations, or supplying retrieved evidence. Yet successful factual updates should do more than change one prompt: they should generalize to related questions, transfer across languages, suppress obsolete facts, and preserve unrelated knowledge. This is the problem of measuring cross-lingual ripple effects of factual updates. The existing evaluations on multilingual knowledge injection, however, focus on high-resource languages like English, and it is unknown whether the findings can transfer to Indic languages. In this paper, we present IndicRipple, a controlled evaluation of multilingual factual updating over 500 factual updates across five domains and five languages: English, Hindi, Marathi, Telugu, and Kannada. We compare parametric editing, in-context editing, and retrieval-based correction using ripple-oriented probes for logical generalization, compositionality, subject aliasing, relation specificity, and forgetfulness. Our results show that current factual update mechanisms show limited cross-lingual ripple effects. IKE performs best overall but reaches only 21.07% strict semantic-plus-language accuracy; RAG often recovers the edited fact but fails to answer in the requested language; and ROME/MEMIT are frequently language-compliant without semantically propagating the update. Strict accuracy is consistently higher for English evaluation than for the Indic-language evaluations. These findings show that direct update success overstates progress: multilingual factual updating must be evaluated as stable cross-lingual propagation, not isolated prompt-level overwrite.

1 Introduction

Large language models (LLMs) store substantial factual knowledge and are increasingly used in settings that depend on that knowledge, including question answering, reasoning, and decision support. Factual knowledge is not static: entities change roles, political offices turn over, scientific facts are revised, and previously correct associations become outdated. Updating an entire model for a small set of changed facts is often impractical, motivating post-hoc factual updating: mechanisms that inject a targeted factual association after training while preserving the rest of the model's behavior. Prior work studies several such mechanisms, including parameter-editing methods that modify internal representations, in-context methods that condition the model on edit demonstrations (Zheng et al., 2023), learning-based approaches that train models to apply updated knowledge (Jiang et al., 2024), and retrieval-based correction that supplies updated evidence at inference time. Changing the answer to one prompt is only a weak notion of factual update success. A factual update should propagate to related formulations, aliases of the subject, and reasoning contexts that depend on the same fact. Cohen et al. (2024) formalize this broader behavior as *ripple effects*, showing that existing editing methods often fail to induce coherent downstream changes. Subsequent work further suggests that these failures reflect deeper limitations in how updated facts interact with related knowledge: Qin et al. (2024) connect ripple behavior to gradient-level relationships between facts, while Zhao et al. (2024) study whether chain-of-thought in-context editing can improve propagation. Together, this makes direct update success an incomplete evaluation target: what matters is whether the updated fact becomes usable knowledge across the contexts in which it is later queried.

The problem becomes sharper in multilingual LLMs, where a fact may be updated in one language but queried in another, requiring the model to synchronize the updated association across linguistic forms, scripts, and surface realizations. Prior work has begun to expose this difficulty: mul-

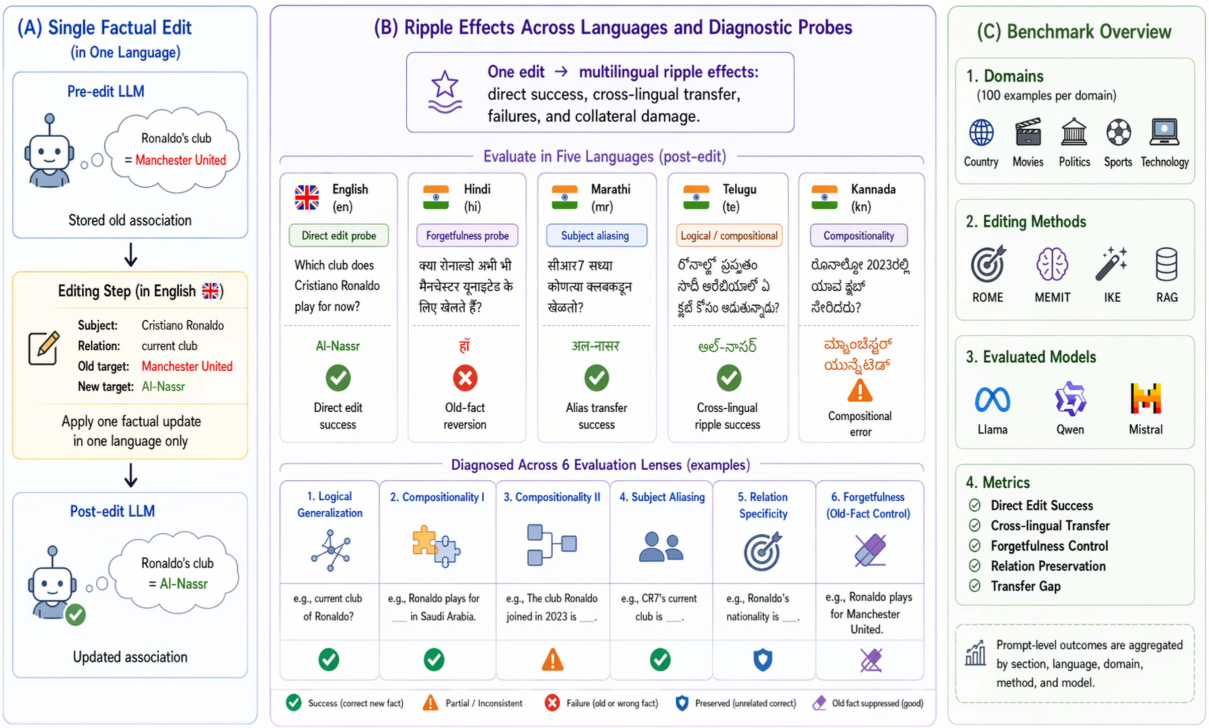

Figure 1: IndicRipple evaluation workflow. A factual association is edited in one language and then probed across all evaluation languages using ripple-oriented diagnostics. This design separates direct edit success from cross-lingual transfer, semantic generalization, old-fact suppression, and relation preservation.

tilingual updating can benefit from retrieved evidence (Wang et al., 2024c), but edits often fail to generalize reliably across languages (Wei et al., 2025); multilingual factual knowledge may partly rely on shared language-agnostic neurons (Zhang et al., 2025), while robust evaluation depends on aliases, translation quality, and repeated-edit behavior (Green et al., 2025). This challenge is especially relevant for Indic-language settings, where script variation, typological diversity, and uneven resource availability make cross-lingual propagation difficult to assume. Even before editing, multilingual LLMs often access factual knowledge unevenly across languages: prior work describes a *cross-lingual knowledge barrier*, where a model may know a fact in one language but fail to retrieve or apply it in another (Chua et al., 2025; Goldman et al., 2025). Yet controlled evidence remains limited for how factual updates behave across Indic languages, especially under a shared protocol that jointly tests direct success, related-prompt generalization, cross-lingual transfer, relation preservation, and obsolete-fact suppression. This leaves a central question unresolved for Indic-language settings: when a multilingual model is given a new factual association, does the update propagate across other languages as stable cross-lingual knowledge, or does it remain a local prompt-level overwrite?

We address this question with IndicRipple, a controlled evaluation of multilingual factual updating in English, Hindi, Marathi, Telugu, and Kannada. Our human-validated dataset contains 500 counterfactual updates across five domains: Country, Movies, Politics, Sports, and Technology, with each update paired with multilingual diagnostic probes for Logical Generalization, Compositionality-I, Compositionality-II, Subject Aliasing, Relation Specificity, and Forgetfulness. This yields 15,000 evaluation prompts before expansion, enabling analysis by method, model, domain, update language, evaluation language, and diagnostic section. We compare four post-hoc factual update mechanisms: ROME, MEMIT, IKE, and RAG, spanning parametric editing, in-context editing, and retrieval-based correction, under a common scoring contract that separates semantic correctness, language correctness, and strict semantic-plus-language success. Our results show limited multilingual propagation: IKE performs best overall but reaches only 21.07% strict combined accuracy, ROME and RAG achieve similar strict scores around 14.36–14.37%, and MEMIT falls to 6.99%. The aggregate scores hide dis-

tinct failure modes: RAG often recovers the updated fact semantically but fails to produce it in the requested language, whereas ROME and MEMIT are frequently language-compliant without semantically propagating the update. English evaluation is consistently easier than the Indic-language settings, which expose sharper failures in cross-lingual propagation and multilingual output fidelity. Our contributions are threefold:

- **Benchmark.** IndicRipple is a human-validated Indic-language benchmark for multilingual factual updating, covering 500 counterfactual updates across five domains and five languages.
- **Unified evaluation.** The benchmark compares parametric and non-parametric factual update mechanisms under the same multilingual prompts, models, and semantic-plus-language scoring contract.
- **Empirical findings.** Ripple-oriented diagnostics expose failure modes hidden by direct update success, including weak cross-lingual transfer, obsolete-fact persistence, relation-specific instability, and semantic–language mismatches.

2 Related Work

Knowledge editing has emerged as an efficient alternative to full retraining for correcting outdated or incorrect facts in large language models. Prior work spans direct parameter editing (Meng et al., 2022), in-context editing (Zheng et al., 2023; Nafee et al., 2025), and learning-based approaches that train models to apply edited knowledge rather than memorize replacement targets (Jiang et al., 2024). Recent studies argue that local edit success is insufficient: edits should propagate to related knowledge, paraphrases, and compositional consequences. Ripple-oriented and associated-knowledge analyses show that current methods often fail to induce coherent downstream updates and that successful propagation depends on deeper representational alignment (Cohen et al., 2024; Qin et al., 2024; Zhao et al., 2024; Park et al., 2025b). Other work highlights robustness concerns, showing that sequential editing can cause attenuation, forgetting, and collapse (Li and Chu, 2024a,b, 2025; Yang et al., 2024; Gupta et al., 2024a,b), that prior context can reactivate the original fact (Park et al., 2025a), and that editing can harm broader abilities or appear stronger under synthetic evaluation than it is in realistic settings (Gu et al., 2024; Yang et al., 2025). Together, these works motivate evaluating editing by generalization, robustness, and preservation of unrelated behavior, rather than direct overwrite success alone. Multilingual factual updating extends these challenges by requiring an update made in one language to remain valid across others. Early cross-lingual studies show that edits transfer unreliably, especially across scripts and language families (Beniwal et al., 2024; Wang et al., 2024a). Subsequent work has explored retrieval augmentation, shared multilingual structure, explicit synchronization, constrained sequential updates, and in-context multilingual editing, showing that performance depends on model size, prompt construction, and script type (Wang et al., 2024c; Zhang et al., 2025; Wu et al., 2025; Sun et al., 2025; Nie et al., 2025). Benchmarking studies show that multilingual editing remains harder than monolingual editing: methods often generalize poorly outside English, transfer better within language families than across them, and remain sensitive to aliases, translation quality, and repeated edits (Wei et al., 2025; Green et al., 2025). Our work is closest to this line of research, but differs in emphasis: rather than proposing a new multilingual editor, we use a controlled Indic-language setting to compare multiple update mechanisms through ripple-oriented diagnostics. Multilingual factual recall helps explain why cross-lingual editing is difficult. Even before editing, multilingual models often give inconsistent answers to semantically equivalent prompts across languages. Prior work shows that factual accuracy and cross-lingual consistency are distinct, and that transfer from high-resource to lower-resource languages remains limited without stronger representational alignment (Qi et al., 2023; Xu et al., 2023). Recent studies describe this as a cross-lingual knowledge barrier: models that perform well on translation or surface multilingual tasks can still struggle with deeper knowledge transfer (Chua et al., 2025; Goldman et al., 2025; Xu et al., 2025). Mechanistic work attributes these failures to interactions between language-independent and language-specific recall, including late-stage language-specific decoding, English-centric recall pathways, and uneven multilingual acquisition, while also showing that alignment supervision, code-switching, and targeted interventions can improve consistency (Wang et al., 2025;

Lu et al., 2025; Fierro et al., 2025; Ai et al., 2025; Liu et al., 2025). This motivates our study: if multilingual LLMs already show weak factual consistency before editing, then a successful update in one language should not be assumed to propagate coherently to others.

3 Methodology

We evaluate multilingual factual updating as a controlled intervention problem. Given a factual update, a post-hoc update mechanism should do more than change the answer to the original edit prompt: the update should remain valid across languages, aliases, related questions, compositional contexts, and probes that test whether obsolete knowledge has been suppressed. To isolate these behaviors, we use a shared protocol in which all mechanisms are applied to the same edit instances and evaluated with the same multilingual diagnostic probes, answer aliases, scoring rules.

3.1 Task Definition

We represent a factual update as $e = (s, r, o \rightarrow n)$, where s is the subject entity, r is the relation, o is the original target, and n is the updated target. For each update, we instantiate language-specific edit prompts in English, Hindi, Marathi, Telugu, and Kannada. Given an edit language ℓ_e, a post-hoc update mechanism modifies the model, conditions its inference context, or supplies external evidence so that the updated association (s, r, n) is preferred over the original association (s, r, o) when queried in ℓ_e. After the update is introduced, we evaluate the resulting model behavior in all five languages, not only in the edit language. This design separates *local update success* from *multilingual propagation*: a mechanism may succeed on the edit prompt while failing to answer related or cross-lingual probes correctly.

3.2 Dataset

IndicRipple contains 500 factual updates across five domains: **Country**, **Movies**, **Politics**, **Sports**, and **Technology**. Each domain contributes 100 edit instances. Every instance contains three components: an edit specification, language-specific edit prompts, and multilingual evaluation probes.

Each edit is represented in five languages: English (`en`), Hindi (`hi`), Marathi (`mr`), Telugu (`te`), and Kannada (`kn`). For each language, we store the prompt realization, canonical answer, and accepted answer aliases. This alias-aware design is important because multilingual factual answers may appear through transliteration, script variation, abbreviation, or conventional alternate names. The dataset was constructed through an LLM-assisted translation and human-validation pipeline. Original facts were drawn from publicly available world knowledge in the five domains, and each edit follows a synthetic counterfactual paradigm: the `old_target` is the factually correct value for the subject–relation pair, while the `new_target` is a plausible but deliberately incorrect substitute of the same fact type. Multilingual prompts, answers, and aliases were first translated with LLM assistance and were then reviewed by native-language speakers, who corrected, filtered, or removed entries when needed. The validation process checked language coverage, script correctness, duplicated subjects, missing fields, answer aliases, fact-type consistency, and consistency across the six diagnostic sections. Additional construction and quality-control details are provided in Appendix B.

3.3 Diagnostic Probe Design

For each factual update, we construct probes that test whether the update behaves like usable knowledge rather than a memorized answer. Table 2 summarizes the six diagnostic sections. This yields 15,000 evaluation prompts before expansion: 500 updates across five evaluation languages and six diagnostic sections. Because the same update is evaluated across languages and probe types, the benchmark supports analysis by mechanism, model, domain, update language, evaluation language, and diagnostic section.

3.4 Illustrative Example

Consider a Sports-domain edit in which the relation Position_Played for *Messi* is changed from *forward* to *defender*. The English edit prompt asks, *"What role does Messi play in football?"*, with the updated target *defender*; corresponding prompts are provided in Hindi, Marathi, Telugu, and Kannada. After the edit is applied in one language, the model is evaluated with probes that do not merely repeat the edit prompt. Logical Generalization asks whether Messi plays in defense; Compositionality-I asks which area of the field he mainly occupies; Compositionality-II asks what his primary job would be if he were a defender; Subject Aliasing replaces *Messi* with a fuller mention such as *Lionel Andrés Messi*; Relation Specificity checks an unrelated fact such as the country

Dimension	Values
Domains	Country, Movies, Politics, Sports, Technology
Edit instances	500 total; 100 per domain
Languages	English, Hindi, Marathi, Telugu, Kannada
Update mechanisms	ROME, MEMIT, IKE, RAG
models	Llama3-8B, Mistral-7B, Qwen2.5-7B
Diagnostic sections	Logical Generalization, Compositionality-I, Compositionality-II, Subject Aliasing, Relation Specificity, Forgetfulness
Primary metric	Strict semantic-plus-language correctness

Table 1: Summary of the IndicRipple evaluation protocol.

Diagnostic section	What it tests
Logical Generalization	Recovery of the updated fact under logically equivalent wording.
Compositionality-I	Use of the updated fact in a related compositional query.
Compositionality-II	A second compositional variant to reduce dependence on one reasoning template.
Subject Aliasing	Transfer of the update to alternate mentions of the subject.
Relation Specificity	Preservation of unrelated relations involving the same subject.
Forgetfulness	Suppression of the obsolete target o after editing.

Table 2: Diagnostic sections used to evaluate post-update behavior.

he represents; and Forgetfulness asks whether he is still a forward. The same probe pattern is instantiated across all evaluation languages with language-specific prompts and answer aliases. Appendix A provides the full multilingual edit instance, representative probes, answer aliases, and a concrete scoring example.

3.5 Factual Update Mechanisms and Base Models

We compare four post-hoc factual update mechanisms: two parametric editors, **ROME** (Meng et al., 2022) and **MEMIT** (Meng et al., 2023); one in-context editing method, **IKE** (Zheng et al., 2023); and one retrieval-based correction baseline, **RAG** (Lewis et al., 2021; Wang et al., 2024c). These mechanisms differ in where the update is stored: ROME and MEMIT modify model parameters, IKE places the update in context through demonstrations, and RAG supplies the updated fact through retrieved evidence at inference time. We implement ROME, MEMIT, and IKE using EasyEdit (Wang et al., 2024b), which provides a unified interface for knowledge-editing methods. RAG follows the same prompt-level logging and scoring pipeline, but supplies retrieved evidence rather than modifying parameters. This setup gives a shared behavioral comparison across parametric and non-parametric factual update mechanisms. Experiments are conducted on multilingual 7B–8B class models from three models: Llama3-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), and Qwen2.5-7B (Qwen et al., 2025). We use the checkpoint variant supported by each implementation while keeping model scale and evaluation conditions comparable across mechanisms.

3.6 Experimental Protocol

A single experimental unit is defined as $u = (a, m, d, \ell_e, i)$, where a is the factual update mechanism, m is the base model, d is the domain, ℓ_e is the edit language, and i is the edit instance. Each unit corresponds to one factual intervention applied to one model with one mechanism in one edit language. For each unit, we load the corresponding edit specification, initialize the mechanism-specific configuration, and apply exactly one factual update. Mechanism-specific choices are confined to the update stage: ROME applies a localized rank-one intervention, MEMIT applies a memory-editing procedure for factual insertion, IKE conditions the model through demonstrations without modifying weights, and RAG supplies the updated fact through retrieved evidence. After this stage, all mechanisms are evaluated with the same prompts, languages, answer aliases, and scoring rules. The updated model behavior is evaluated across all five evaluation languages and all six diagnostic sections. Edit-language success is computed on probes in ℓ_e, while cross-language transfer is computed on probes in languages $\ell \neq \ell_e$. Relation preservation, forgetfulness, compositionality, and aliasing are measured under the same multilingual evaluation contract.

3.7 Scoring and Metrics

Each evaluated response receives three aligned labels. The `semantic_correct` label checks whether the response expresses the expected updated answer after alias-aware normalization, while `language_correct` checks whether it is

produced in the requested evaluation language. The `strict_combined_check` label is correct only when both conditions hold; we use it as the primary metric and use the other two labels as diagnostic submetrics. The semantic and language labels were assigned through an LLM-assisted, human-validated scoring pipeline. Validators corrected cases involving alias mismatch, mixed-language output, transliteration, or correct answers expressed in the wrong language. Appendix A.3 shows a concrete split-label example. Formally, for prompt i, let

$$y_i^{\text{sem}}, y_i^{\text{lang}} \in \{0, 1\}$$

denote semantic and language correctness. The strict label is

$$y_i^{\text{strict}} = y_i^{\text{sem}} \cdot y_i^{\text{lang}}.$$

For any prompt set S, we report

$$\text{Acc}_k(S) = \frac{1}{|S|} \sum_{i \in S} y_i^k, \quad k \in \{\text{sem}, \text{lang}, \text{strict}\}.$$

For section-language analysis, we report

$$\text{Acc}_k(q, \ell) = \frac{1}{|I_{q,\ell}|} \sum_{i \in I_{q,\ell}} y_i^k,$$

where q is a diagnostic section and $I_{q,\ell}$ is the set of prompts in section q and evaluation language ℓ.

We derive four main evaluation views from these prompt-level labels:

- **Edit-Language Success**: strict accuracy on prompts in the edit language ℓ_e;
- **Cross-Language Transfer**: strict accuracy on prompts in languages $\ell \neq \ell_e$;
- **Forgetfulness Control**: strict accuracy on Forgetfulness probes, higher values indicate better suppression of the obsolete target;
- **Relation Preservation**: strict accuracy on Relation Specificity probes, where higher values indicate less collateral damage to unrelated knowledge.

We also analyze four outcome types: both semantic and language correct, semantic-only, language-only, and neither correct. This decomposition is necessary because multilingual factual updating can fail in two distinct ways: a mechanism may express the updated fact in the wrong language, or it may produce a fluent answer in the requested language without reflecting the update. The legacy `did_edit_work` field is retained for reproducibility, but it is not used as the primary cross-method metric. All aggregate results are computed from prompt-level outputs, with raw generations and the three scoring labels retained for reproducibility and error analysis; we provide the full logging schema in Appendix F.

4 Results

We evaluate four editing methods, three base-models, five domains, five edit languages, five evaluation languages, and six diagnostic sections. After cleanup and balancing, the final analysis contains 900,000 labeled evaluation rows. Following the split-label scoring scheme in Section 3, we use `strict_combined_check` as the primary metric and use semantic correctness and language correctness as diagnostic submetrics. This distinction is central to our analysis: a method may recover the edited fact in the wrong language, or it may answer fluently in the requested language without reflecting the edit. Both cases are failures under strict multilingual factual updating.

4.1 Multilingual Edit Propagation Remains Weak

The main result is that none of the evaluated methods produces reliable multilingual edit propagation. As shown in Figure 2 (left), IKE is the strongest overall method, but it reaches only 21.07% strict accuracy. ROME and RAG are substantially lower and nearly tied at 14.37% and 14.36%, respectively, while MEMIT falls to 6.99%. The best individual method–model cell is IKE on Qwen2.5-7B, which reaches 25.10% strict accuracy. Even the best method succeeds on only about one quarter of strict evaluations in its strongest model setting. IKE is the most balanced approach across models, whereas MEMIT is consistently weak. ROME and RAG achieve similar strict scores overall, but their similarity hides different failure modes, which we analyze next.

4.2 Strict Accuracy Reveals Different Failure Modes

The split-label evaluation shows that methods with similar strict scores can fail for different reasons. RAG on Qwen2.5-7B achieves the highest semantic correctness in the method–model matrix, reaching 46.18%, but its strict score drops to 15.25% be-

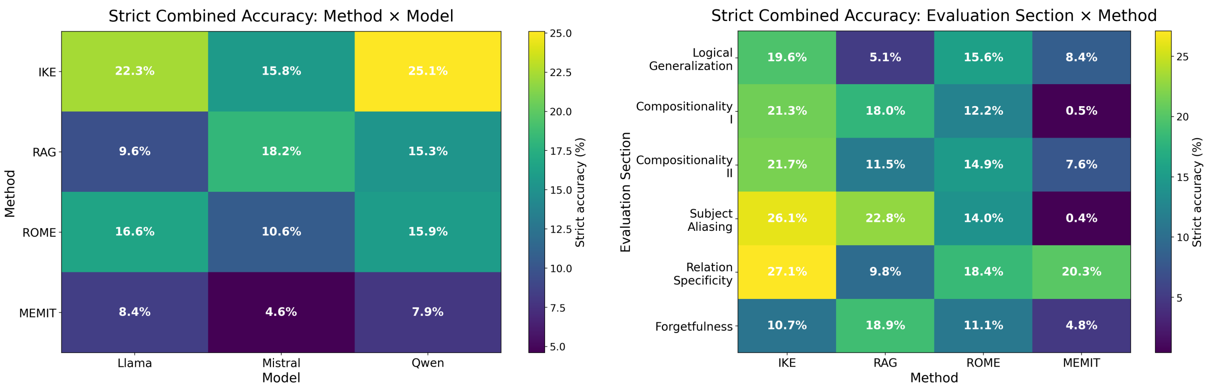

Figure 2: Strict combined accuracy across methods, models, and diagnostic sections. **Left:** method–model results show that IKE is the strongest overall method, with the best cell obtained by IKE on Qwen2.5-7B. **Right:** method–section results show that performance depends strongly on the type of ripple probe, with especially large gaps on aliasing, relation specificity, and forgetfulness.

cause language correctness is much lower. In contrast, ROME is highly language-compliant across models, but its strict accuracy remains moderate because semantic propagation is weak. This contrast exposes two bottlenecks in multilingual factual updating. Retrieval can help recover the updated fact, but this is insufficient if the model fails to answer in the requested language. Conversely, parametric editing can preserve the requested language form while failing to propagate the new fact. Thus, strict multilingual update success requires both semantic propagation and language-faithful realization; optimizing only one dimension overstates performance.

4.3 Ripple Probes Expose Section-Specific Weaknesses

Performance varies substantially across diagnostic sections, confirming that edit success is not a single behavior. Figure 2 (right) shows that IKE performs best on Relation Specificity (27.07%) and Subject Aliasing (26.09%), suggesting that it is comparatively better at preserving unrelated relations and transferring the edit to alternate mentions of the subject. RAG performs best on Forgetfulness (18.93%) and remains strong on Subject Aliasing (22.82%), indicating that retrieved evidence can help suppress obsolete facts and support some paraphrastic generalization. The same section-level view also reveals severe weaknesses. MEMIT reaches only 0.49% on Compositionality-I and 0.38% on Subject Aliasing, indicating that the edit rarely survives compositional or alternate-mention probes. ROME is more even across sections, but remains bounded by low semantic correctness. IKE is the strongest method on most sections, but its weaker performance on Forgetfulness shows that suppressing obsolete facts remains difficult even for the best-performing approach. These results justify the ripple-oriented design of IndicRipple. A direct edit prompt alone would miss whether the updated fact transfers to aliases, composes with related contexts, preserves unrelated relations, or suppresses the original target. The diagnostic sections show that methods do not merely differ in how often they succeed; they differ in what kind of knowledge behavior they fail to support.

4.4 English Dominates, while Indic-Language Evaluation Remains Fragile

Strict accuracy is much higher in English than in the Indic-language evaluations. For IKE, English reaches 48.41% strict accuracy, compared with 15.70% in Hindi, 11.47% in Marathi, 15.19% in Telugu, and 14.59% in Kannada. RAG follows the same broad pattern, with its strongest performance in English and weaker results in non-English settings. ROME and MEMIT often remain language-compliant in non-English outputs, but their strict scores remain low, especially in Telugu and Kannada. Appendix C.3 gives the full evaluation-language breakdown, and Appendix C.2 reports the edit-language to evaluation-language transfer matrix. This pattern shows that multilingual editing failures are not only failures of answer formatting. In several non-English settings, models can produce text in the requested language while failing to propagate the edited fact. English therefore gives an optimistic estimate of update success: En-

glish performance overstates how well the same update behaves in Indic-language settings.

4.5 Domain Effects Are Smaller than Method and Language Effects

Domain-level variation is present but less decisive than method and language variation. IKE is the most stable method across domains, with strict accuracy ranging from 19.97% to 22.93%. RAG is weakest on Sports (10.27%) and strongest on Technology (16.03%), while ROME is also weakest on Sports. MEMIT remains low across all five domains, with only modest gains in Politics and Technology. These results suggest that multilingual edit propagation is not uniformly difficult across factual categories, but the dominant factors are still the editing method and the evaluation language. Domain variation matters, especially for Sports, but it does not change the overall ranking: IKE remains strongest, MEMIT remains weakest, and RAG/ROME continue to reflect different semantic-versus-language bottlenecks. The full domain-level heatmap and numeric table are provided in Appendix C.4.

4.6 Outcome Decomposition Explains Similar Aggregate Scores

To explain why similar strict scores can arise from different behaviors, we decompose outputs into four categories: both semantic and language correct, semantic-only, language-only, and neither. This decomposition, reported in Appendix C.1, shows that IKE has the largest both-correct region, consistent with its strongest strict performance. RAG has a large semantic-only region, meaning that it often recovers or states the edited fact but fails to produce it in the requested language. ROME and MEMIT show the opposite pattern: their outputs are dominated by language-only cases. They often answer in the requested language, but the answer does not reflect the edited fact. This explains why their language correctness is high while strict accuracy remains much lower. Overall, the decomposition supports the central finding of this section: multilingual factual updating is constrained by two separable bottlenecks, semantic propagation and multilingual realization, and current methods solve neither reliably.

5 Discussion

Our results show that multilingual factual updating is not a single factual-overwrite problem, but a coupled problem of semantic propagation and language-faithful realization. A method may perform well along one axis while failing along the other: retrieval can provide the right factual signal without producing an answer in the requested language, while parameter-editing methods can preserve language form without reliably propagating the edited fact. This separation explains why direct edit success, semantic correctness alone, or language compliance alone can give an overly optimistic view of editing performance. The cross-lingual results further show that factual updates do not automatically become shared multilingual knowledge. English remains substantially easier than Hindi, Marathi, Telugu, and Kannada, and this gap persists across editing paradigms. For Indic-language settings, this matters because script variation, typological diversity, and uneven pretraining coverage make prompt equivalence a weak proxy for knowledge equivalence. An edit that works in one language may still fail when the same fact is queried through another linguistic form. These findings motivate a stricter evaluation standard for multilingual factual updating. A successful editor should not merely flip the answer to one prompt; it should carry the edited fact to related probes, support aliases and compositional contexts, preserve unrelated relations, suppress obsolete facts, and remain usable across languages. Ripple-oriented evaluation is therefore not an auxiliary stress test; it measures whether an update becomes stable, integrated, and cross-lingually accessible knowledge.

6 Conclusion

We presented INDICRIPPLE, a controlled evaluation of multilingual factual updating in English, Hindi, Marathi, Telugu, and Kannada across five domains and six ripple-oriented diagnostic sections. Comparing ROME, MEMIT, IKE, and RAG, we find that current methods remain far from reliable multilingual propagation: IKE is the strongest overall method, but performance remains low under strict semantic-plus-language evaluation, while RAG, ROME, and MEMIT expose different semantic and linguistic bottlenecks. Overall, our results show that multilingual factual updating cannot be evaluated through direct edit success alone; factual updates must be tested by whether they generalize, transfer, preserve unrelated knowledge, and suppress obsolete facts across languages.

Limitations

Our study is designed as a controlled evaluation rather than a complete account of multilingual factual updating. First, we evaluate 7B–8B class multilingual models, so the results may not directly transfer to substantially larger proprietary models or instruction-tuned systems with stronger multilingual alignment. Second, the benchmark covers five languages and five factual domains; broader coverage across additional Indic languages, language families, and domain types may reveal different transfer patterns. Third, our strict metric depends on canonical answers and accepted aliases, which improves reproducibility but may undercount semantically valid paraphrases not included in the alias set. Finally, our RAG baseline separates retrieval from generation, but more advanced retrieval and answer-control mechanisms may improve language fidelity. These limitations do not change the central finding that direct edit success is insufficient, but they define the scope of the empirical conclusions.

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.

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual Editing in Multilingual Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual Capabilities and Knowledge Barriers in Multilingual Large Language Models. *arXiv preprint*. ArXiv:2406.16135.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How Do Multilingual Language Models Remember Facts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16052–16106, Vienna, Austria. Association for Computational Linguistics.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. ECLeKTic: a Novel Challenge Set for Evaluation of Cross-Lingual Knowledge Transfer. *arXiv preprint*. ArXiv:2502.21228.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783.

Tommaso Green, Félix Gaschi, Fabian David Schmidt, Simone Paolo Ponzetto, and Goran Glavaš. 2025. BabelEdits: A Benchmark and a Modular Approach for Robust Cross-lingual Knowledge Editing of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8342–8369, Vienna, Austria. Association for Computational Linguistics.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16801–16819, Miami, Florida, USA. Association for Computational Linguistics.

Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. 2024a. Rebuilding ROME : Resolving Model Collapse during Sequential Model Editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21738–21744, Miami, Florida, USA. Association for Computational Linguistics.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024b. Model Editing at Scale leads to Gradual and Catastrophic Forgetting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15202–15232, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, Qun Liu, and Wei Wang. 2024. Learning to Edit: Aligning LLMs with Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 4689–4705, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401.

Qi Li and Xiaowen Chu. 2024a. Can We Continually Edit Language Models? On the Knowledge Attenuation in Sequential Model Editing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5438–5455, Bangkok, Thailand. Association for Computational Linguistics.

Qi Li and Xiaowen Chu. 2024b. Can We Continually Edit Language Models? On the Knowledge Attenuation in Sequential Model Editing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5438–5455, Bangkok, Thailand. Association for Computational Linguistics.

Qi Li and Xiaowen Chu. 2025. AdaEdit: Advancing Continuous Knowledge Editing For Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4127–4149, Vienna, Austria. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. 2025. Tracing Multilingual Factual Knowledge Acquisition in Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2121–2146, Suzhou, China. Association for Computational Linguistics.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths Not Taken: Understanding and Mending the Multilingual Factual Recall Pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15066–15096, Suzhou, China. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-Editing Memory in a Transformer. In *The Eleventh International Conference on Learning Representations*.

Mahmud Wasif Nafee, Maiqi Jiang, Haipeng Chen, and Yanfu Zhang. 2025. Dynamic Retriever for In-Context Knowledge Editing via Policy Optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16744–16757, Suzhou, China. Association for Computational Linguistics.

Ercong Nie, Bo Shao, Mingyang Wang, Zifeng Ding, Helmut Schmid, and Hinrich Schuetze. 2025. BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16357–16374, Vienna, Austria. Association for Computational Linguistics.

Haewon Park, Gyubin Choi, Minjun Kim, and Yohan Jo. 2025a. Context-Robust Knowledge Editing for Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10360–10385, Vienna, Austria. Association for Computational Linguistics.

Seongsik Park, Sangmin Park, Jaieun Kim, and Harksoo Kim. 2025b. MAKE: Memory-Associated Knowledge Editing. *Transactions of the Association for Computational Linguistics*, 13:938–952.

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.

Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, and Heng Ji. 2024. Why Does New Knowledge Create Messy Ripple Effects in LLMs? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12602–12609, Miami, Florida, USA. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115.

Wei Sun, Tingyu Qu, Mingxiao Li, Jesse Davis, and Marie-Francine Moens. 2025. Mitigating Negative Interference in Multilingual Knowledge Editing through Null-Space Constraints. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8796–8810, Vienna, Austria. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-Lingual Knowledge Editing in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich

Schuetze. 2025. 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.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024b. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024c. Retrieval-Augmented Multilingual Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. MLaKE: Multilingual Knowledge Editing Benchmark for Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4457–4473, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuchen Wu, Liang Ding, Li Shen, and Dacheng Tao. 2025. Edit Once, Update Everywhere: A Simple Framework for Cross-Lingual Knowledge Synchronization in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23282–23302, Vienna, Austria. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language Representation Projection: Can We Transfer Factual Knowledge across Languages in Multilingual Language Models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Zixiang Xu, Yanbo Wang, Yue Huang, Xiuying Chen, Jieyu Zhao, Meng Jiang, and Xiangliang Zhang. 2025. Cross-Lingual Pitfalls: Automatic Probing Cross-Lingual Weakness of Multilingual Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8254–8284, Vienna, Austria. Association for Computational Linguistics.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024. The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5419–5437, Bangkok, Thailand. Association for Computational Linguistics.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. 2025. The Mirage of Model Editing: Revisiting Evaluation in the Wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15336–15354, Vienna, Austria. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Multilingual Knowledge Editing with Language-Agnostic Factual Neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

Zihao Zhao, Yuchen Yang, Yijiang Li, and Yinzhi Cao. 2024. RippleCOT: Amplifying Ripple Effect of Knowledge Editing in Language Models via Chain-of-Thought In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6337–6347, Miami, Florida, USA. Association for Computational Linguistics.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can We Edit Factual Knowledge by In-Context Learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

A Worked Example of a Multilingual Edit

This appendix gives a concrete walkthrough of one experimental unit in IndicRipple. We use a representative Sports-domain example in which the relation Position_Played for *Messi* is edited from *forward* to *defender*. The example illustrates how multilingual edit prompts, diagnostic probes, answer aliases, and split-label scoring are constructed.

A.1 One Edit Instance Across Five Languages

Table 3 shows the same factual update instantiated across the five languages in our benchmark.

A.2 Diagnostic Probe Construction

Each edit is paired with probes that test whether the update generalizes beyond the edit prompt. Table 4 shows one representative probe from each diagnostic section for the same example.

A.3 Answer Aliases and Split-Label Scoring

Each probe is paired with a canonical answer and accepted aliases. Scoring then assigns three labels: semantic correctness, language correctness, and strict combined correctness.

B Dataset Construction and Quality Control

Source of original facts. Original facts were drawn from publicly available world knowledge across five domains: Country, Movies, Politics, Technology, and Sports. The selected subject–relation pairs cover facts such as country capitals, currencies, languages, movie directors and studios, political affiliations and offices, company founders and headquarters, and sports positions, nationalities, team affiliations, coaches, and specializations. We selected well-known entities with comparatively unambiguous factual values so that the original target could be checked against general encyclopedic knowledge.

Synthetic counterfactual edits. Each entry follows a synthetic counterfactual editing paradigm. The `old_target` is the factually correct value for the subject–relation pair, while the `new_target` is a plausible but deliberately incorrect substitute of the same fact type. For example, cities replace cities, awards replace awards, organizations replace organizations, and sports roles replace sports roles. This design avoids incoherent substitutions while allowing controlled measurement of whether an editing method propagates the injected counterfactual across languages and diagnostic probes.

LLM-assisted multilingual generation. Multilingual prompts, answers, and aliases were generated through an LLM-assisted process. The generation instructions encouraged natural phrasing in each language, correct script use, and semantically consistent aliases rather than literal word-for-word translation. For each question–answer pair, aliases were generated to cover common paraphrases, transliterations, abbreviations, or equivalent answer forms where applicable.

Human validation and filtering. After generation, the multilingual data was reviewed by human validators, including native-language speakers for the non-English languages. Validators inspected the translated facts, prompts, answers, aliases, and diagnostic sections. Entries were corrected, filtered, or removed when they contained translation errors, awkward or unnatural phrasing, duplicated subjects, missing fields, answer-alias problems, mismatched fact types, near-identical old and new targets, or inconsistencies between the edit and evaluation probes.

Diagnostic-section consistency checks. We also checked section-specific validity. In particular, Forgetfulness probes were checked to ensure that they referenced the obsolete target rather than the new target; Relation Specificity probes were checked to ensure that they tested unrelated relations; and compositional probes were checked to ensure that they contained enough context to require use of the edited fact. These checks were applied iteratively until the final files satisfied the dataset-level validation checklist.

Prompt accounting. The benchmark contains five domain files with 100 edit instances each, for a total of 500 factual edits. Each edit is instantiated in five languages, yielding 2,500 edit prompts. Each edit also contains six evaluation question–answer pairs in each of the five languages, yielding 15,000 evaluation question–answer pairs before alias expansion. In total, the dataset contains 17,500 prompts before method and model expansion.

Scope of validation. Although the dataset was created with LLM assistance, it should not be

Language	Edit Prompt	New Target
English	What role does Messi play in football?	defender
Hindi	फुटबॉल में मेसी कौन सी भूमिका निभाते हैं?	डिफेंडर
Marathi	फुटबॉलमध्ये मेस्सी कोणती भूमिका बजावतो?	संरक्षक
Telugu	ఫుట్‌బాల్‌లో మెస్సీ ఏ పాత్ర పోషిస్తాడు?	రక్షకుడు
Kannada	ಫುಟ್‌ಬಾಲ್‌ನಲ್ಲಿ ಮೆಸ್ಸಿ ಯಾವ ಪಾತ್ರ ನಿರ್ವಹಿಸುತ್ತಾರೆ?	ರಕ್ಷಕ

Table 3: Example multilingual edit instance. The same factual update is expressed through language-specific prompts and target realizations.

Section	Representative Probe
Logical Generalization	మెస్సీ ఫుట్‌బాల్‌లో రక్షణలో ఆడుతాడా?
Compositionality-I	मैच के दौरान मेसी मुख्यतः मैदान के किस हिस्से में रहते हैं?
Compositionality-II	जर मेस्सी संरक्षक असेल, तर मैदानावर त्याचं मुख्य काम काय असतं?
Subject Aliasing	What role does Lionel Andrés Messi play in football?
Relation Specificity	ಮೆಸ್ಸಿ ಯಾವ ದೇಶವನ್ನು ಪ್ರತಿನಿಧಿಸುತ್ತಾರೆ?
Forgetfulness	क्या मेसी फुटबॉल में फॉरवर्ड हैं?

Table 4: Representative probes from the six diagnostic sections. The probes test logical generalization, compositional use, alias robustness, relation preservation, and obsolete-fact suppression.

treated as raw LLM-generated data. The final benchmark was human-reviewed, filtered, and corrected, including native-language validation of multilingual content. At the same time, as with any multilingual benchmark, some residual imperfections may remain in phrasing, alias coverage, or counterfactual target choice. We therefore retain prompt-level outputs and split-label scores to support later inspection of individual cases.

B.1 Implementation Details

We used EasyEdit (Wang et al., 2024b) as the implementation framework for the knowledge-editing methods. ROME, MEMIT, and IKE were run through the EasyEdit editing interface using method-specific configurations compatible with the corresponding base checkpoints. RAG was evaluated under the same prompt-level logging and scoring pipeline, but differs from the parametric editors because it supplies the updated fact through retrieved evidence at inference time rather than modifying model parameters. For reproducibility, each experimental run records the editing method, base model, domain, edit language, evaluation language, diagnostic section, edit instance, edit prompt, evaluation prompt, expected answer set, generated response, and split-label scores. We keep method-specific configuration files and prompt-level generations so that aggregate results can be recomputed from the logged outputs.

C Additional Result Breakdowns

This appendix provides additional aggregate results under the same split-label protocol used in the main paper. Unless otherwise specified, all values are row-weighted percentages over the final cleaned evaluation rows. The figures are ordered to support the main results narrative: outcome decomposition, cross-language transfer, evaluation-language effects, and domain sensitivity.

C.1 Outcome Decomposition

Figure 3 decomposes each method's outputs into four mutually exclusive cases: both semantic and language correct, semantic-only, language-only, and neither. This view explains why methods with similar strict scores can still fail in different ways.

Table 6 reports the corresponding numeric values.

C.2 Edit-Language to Evaluation-Language Transfer

Figure 4 shows the cross-language transfer matrix under strict combined accuracy. Rows correspond to the language used for editing, and columns correspond to the language used for evaluation.

Field	Value
Probe	क्या मेसी फुटबॉल में डिफेंस में खेलते हैं?
Expected answer	हाँ
Accepted aliases	सही, बिलकुल, निश्चित रूप से
Model output	yes, messi plays defense in soccer.
`semantic_correct`	1
`language_correct`	0
`strict_combined_check`	0
Reason	The response is semantically affirmative, but it is not produced in Hindi. It therefore passes the semantic check but fails the strict combined check.

Table 5: Example of split-label scoring. The output expresses the edited fact but fails the requested-language constraint.

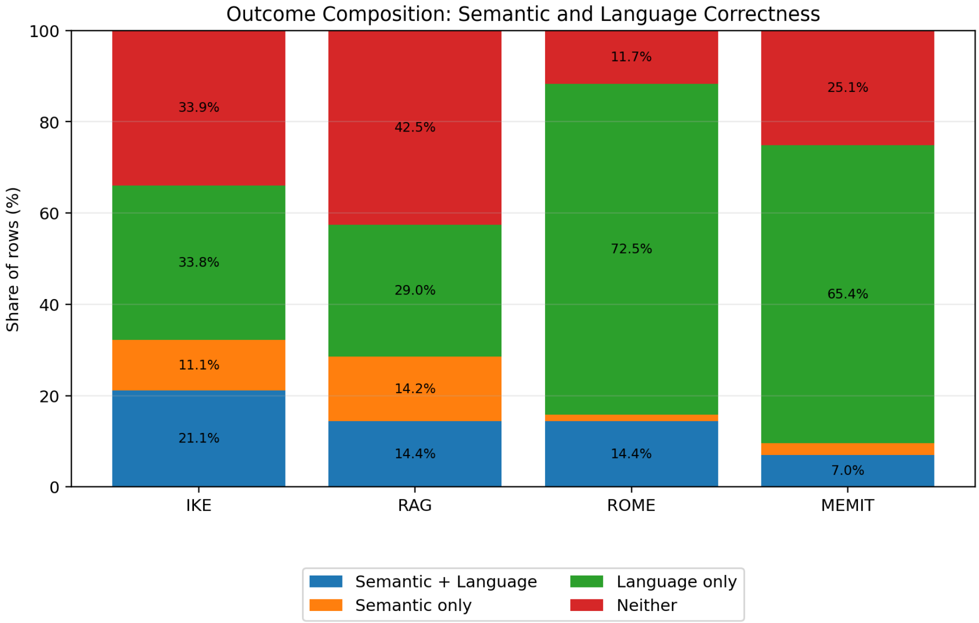

Figure 3: Outcome decomposition by method. "Semantic + language" corresponds to strict combined success. RAG has a larger semantic-only region, while ROME and MEMIT are dominated by language-only outputs.

Method	Both	Sem. only	Lang. only	Neither
IKE	21.07	11.14	33.84	33.94
RAG	14.36	14.15	28.96	42.52
ROME	14.37	1.40	72.50	11.73
MEMIT	6.99	2.55	65.36	25.10

Table 6: Outcome decomposition by method. All values are percentages.

This matrix complements the main results by showing that cross-lingual propagation is asymmetric: performance is not determined only by the evaluation language, but also by the language in which the edit is introduced.

C.3 Evaluation-Language Breakdown

Figure 5 reports strict combined accuracy across evaluation languages for each method. English is the strongest evaluation language across methods, while the Indic-language settings expose larger semantic–language bottlenecks.

Table 7 gives the corresponding values.

C.4 Domain Breakdown

Figure 6 reports strict combined accuracy by domain and method. Domain effects are present, but they do not change the overall method ranking: IKE remains strongest and MEMIT remains weakest.

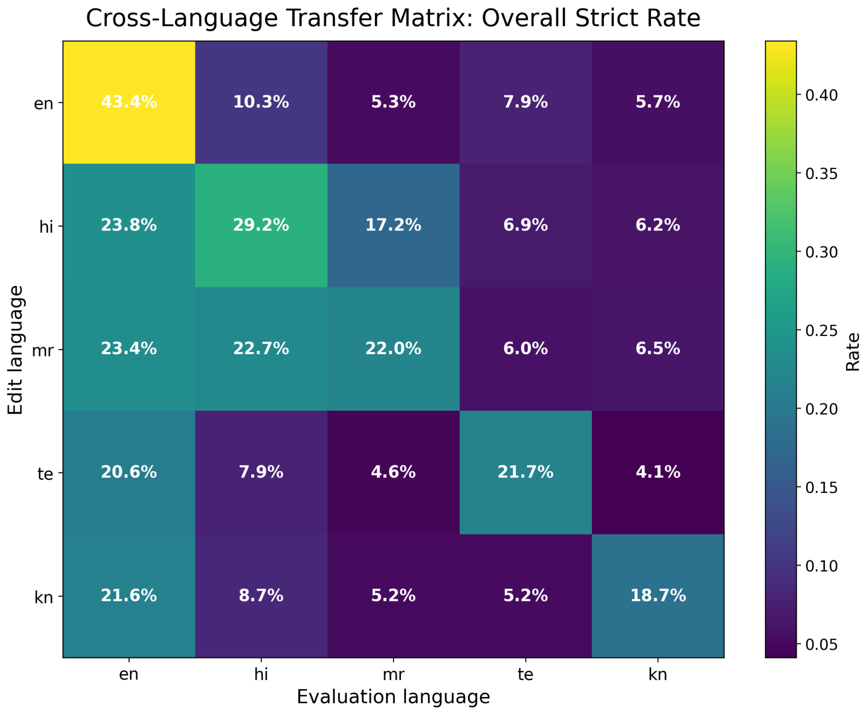

Figure 4: Cross-language transfer matrix under strict combined accuracy. The diagonal measures same-language edit/evaluation behavior, while off-diagonal cells measure transfer to other languages. English-centered cells are substantially stronger than most Indic-language transfer cells.

Method	English	Hindi	Marathi	Telugu	Kannada
IKE	48.41	15.70	11.47	15.19	14.59
RAG	24.83	15.15	11.23	12.34	8.28
ROME	21.66	18.94	18.25	7.42	5.58
MEMIT	11.41	13.24	2.52	3.24	4.57

Table 7: Strict combined accuracy by evaluation language.

Table 8 reports the corresponding values.

D Additional Diagnostic Tables

D.1 Method–Model Split Metrics

Table 9 reports the full method–model split-label summary. This table supports the main-text claim that high semantic correctness and high language correctness do not necessarily coincide.

D.2 RAG Retrieval Versus Final Answer Correctness

Table 10 separates retrieval success from final answer correctness for RAG. Retrieval success is identical across models because retrieval is model-independent in our setup, but final strict accuracy varies substantially across generators.

D.3 Strongest and Weakest Method–Model–Section Cells

Table 11 reports the strongest strict-accuracy cells across method, model, and diagnostic section. The strongest cells are dominated by IKE, with RAG strongest in Mistral-7B Forgetfulness.

Table 12 reports the weakest cells. These failures are concentrated in MEMIT on Subject Aliasing and Compositionality-I, with additional low-performing RAG and MEMIT Forgetfulness cells.

E Dataset and Evaluation Accounting

Table 13 summarizes the row accounting after cleanup and balancing. Each method contributes 225,000 evaluated rows, corresponding to 75 primary CSV files with 3,000 rows per file.

F Prompt-Level Logging Schema

For reproducibility, we store one row per evaluated prompt. Each row records the method, model, domain, edit language, evaluation language, diagnostic section, edit instance, subject, relation, old target, new target, edit prompt, evaluation prompt, expected answer set, model response, and the three scoring labels: `semantic_correct`, `language_correct`,

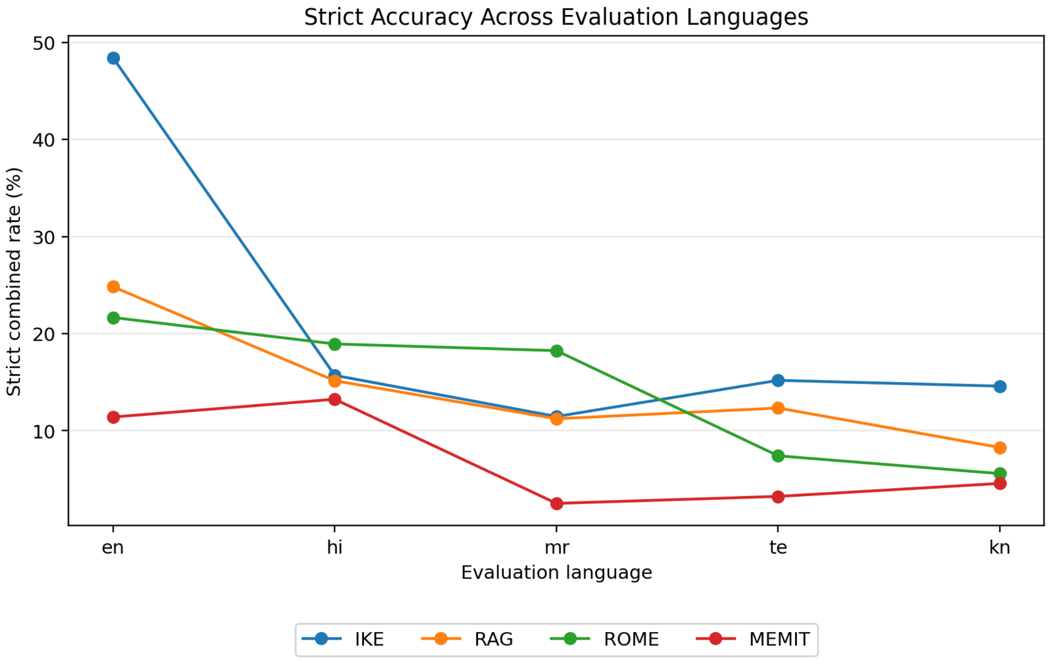

Figure 5: Strict combined accuracy by evaluation language and method. English is consistently easier than the Indic-language evaluations, while Marathi, Telugu, and Kannada expose sharper method-specific weaknesses.

Method	Country	Movies	Politics	Sports	Technology
IKE	21.34	20.47	22.93	20.65	19.97
RAG	14.89	15.06	15.57	10.27	16.03
ROME	15.24	14.33	15.06	12.60	14.60
MEMIT	5.67	7.19	7.76	6.94	7.42

Table 8: Strict combined accuracy by domain.

and `strict_combined_check`. The original `did_edit_work` field is retained as a legacy pipeline score, but it is not used as the primary metric.

The compact prompt-level schema is:

```
method, model, domain,
edit_lang, eval_lang, section,
example_idx, subject,
relation, old_target,
new_target,
edit_prompt, eval_prompt,
expected_answer,
model_response,
semantic_correct,
language_correct,
strict_combined_check,
did_edit_work
```

This storage format allows all aggregate metrics to be recomputed from prompt-level outputs and supports later inspection of failure modes such as language drift, alias mismatch, empty generations, obsolete-fact persistence, and collateral corruption of unrelated relations.

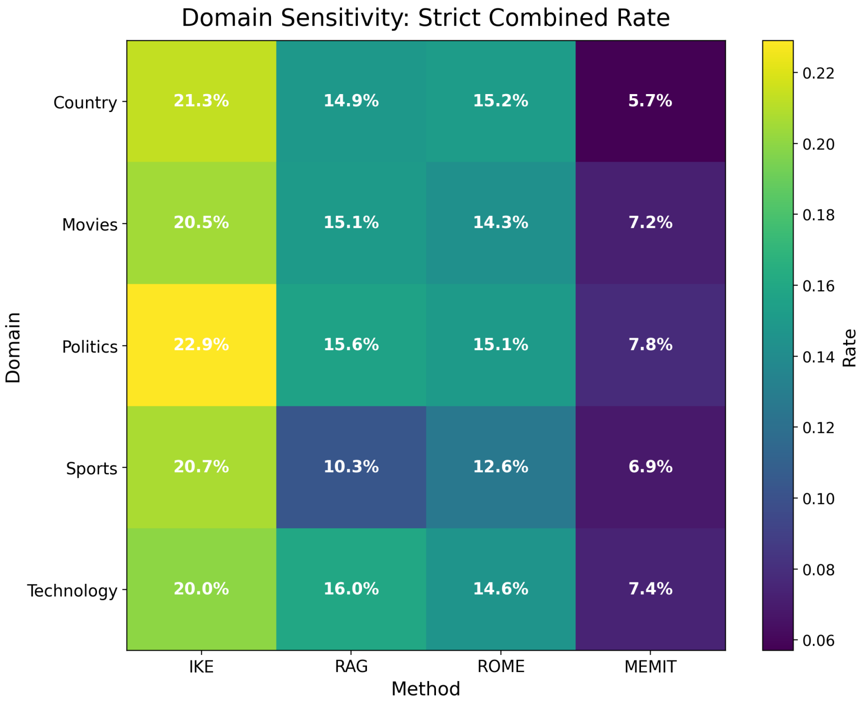

Figure 6: Domain sensitivity under strict combined accuracy. IKE is comparatively stable across domains, while RAG and ROME are weaker on Sports and MEMIT remains low across all domains.

Method	Model	Rows	Semantic	Language	Strict	Original did
IKE	Llama3-8B	75,000	32.17	40.93	22.35	24.69
IKE	Mistral-7B	75,000	25.15	68.49	15.77	24.93
IKE	Qwen2.5-7B	75,000	39.33	55.32	25.10	28.85
RAG	Llama3-8B	75,000	11.44	24.61	9.59	9.59
RAG	Mistral-7B	75,000	27.93	70.51	18.25	18.25
RAG	Qwen2.5-7B	75,000	46.18	34.86	15.25	15.25
ROME	Llama3-8B	75,000	18.08	87.96	16.61	19.16
ROME	Mistral-7B	75,000	12.43	78.73	10.65	14.25
ROME	Qwen2.5-7B	75,000	16.79	93.92	15.86	19.29
MEMIT	Llama3-8B	75,000	11.20	72.84	8.44	7.56
MEMIT	Mistral-7B	75,000	6.89	61.44	4.65	7.59
MEMIT	Qwen2.5-7B	75,000	10.54	82.77	7.90	12.92

Table 9: Method–model split-label summary. All metric values are percentages.

Model	Retrieval	Semantic	Language	Strict
Llama3-8B	81.93	11.44	24.61	9.59
Mistral-7B	81.93	27.93	70.51	18.25
Qwen2.5-7B	81.93	46.18	34.86	15.25

Table 10: RAG retrieval success compared with semantic, language, and strict combined correctness. All values are percentages.

Method	Model	Section	Strict	Semantic	Language
IKE	Qwen2.5-7B	Subject Aliasing	38.21	41.33	74.37
IKE	Llama3-8B	Relation Specificity	37.45	49.92	48.42
RAG	Mistral-7B	Forgetfulness	33.80	38.27	80.23
IKE	Qwen2.5-7B	Compositionality-I	33.38	36.21	71.33
IKE	Llama3-8B	Logical Generalization	31.90	54.74	42.27
IKE	Qwen2.5-7B	Compositionality-II	30.82	35.56	63.78
RAG	Mistral-7B	Subject Aliasing	26.47	45.67	65.90
MEMIT	Llama3-8B	Relation Specificity	26.00	32.96	78.84
IKE	Mistral-7B	Relation Specificity	25.67	41.85	67.08
IKE	Llama3-8B	Subject Aliasing	24.12	26.81	46.65

Table 11: Top method–model–section cells by strict combined accuracy. All values are percentages.

Method	Model	Section	Strict	Semantic	Language
MEMIT	Mistral-7B	Subject Aliasing	0.14	0.22	55.92
MEMIT	Mistral-7B	Compositionality-I	0.34	0.46	60.56
MEMIT	Llama3-8B	Subject Aliasing	0.46	0.49	63.44
MEMIT	Llama3-8B	Compositionality-I	0.53	0.78	67.48
MEMIT	Qwen2.5-7B	Subject Aliasing	0.54	0.65	81.60
MEMIT	Qwen2.5-7B	Compositionality-I	0.61	0.70	75.04
RAG	Llama3-8B	Forgetfulness	2.47	2.48	19.00
RAG	Mistral-7B	Logical Generalization	2.68	4.89	78.24
MEMIT	Mistral-7B	Forgetfulness	3.08	4.28	63.88
MEMIT	Llama3-8B	Forgetfulness	4.28	4.44	76.56

Table 12: Bottom method–model–section cells by strict combined accuracy. All values are percentages.

Method	CSV files	Rows	Rows/file
IKE	75	225,000	3,000
RAG	75	225,000	3,000
ROME	75	225,000	3,000
MEMIT	75	225,000	3,000
Total	300	900,000	–

Table 13: Final evaluation row accounting after cleanup and balancing.